Oversampling PCM Techniques and Optimum Noise Shapers for Quantizing a Class of Nonbandlimited Signals

Jamal Tuqan, *Member, IEEE* and P. P. Vaidyanathan, *Fellow, IEEE*

***Abstract*—We consider the efficient quantization of a class of *nonbandlimited* signals, namely, the class of discrete-time signals that can be recovered from their decimated version. The signals of interest are modeled as the output of a single FIR interpolation filter (single band model) or, more generally, as the sum of the outputs of L FIR interpolation filters (multiband model). By definition, these *nonbandlimited* signals are oversampled, and it is therefore reasonable to expect that we can reap the same benefits of well-known efficient A/D techniques that apply only to bandlimited signals. Indeed, by using appropriate multirate models and reconstruction schemes, we first show that we can obtain a great reduction in the quantization noise variance due to the oversampled nature of the signals. We also show that we can achieve a substantial decrease in bit rate by appropriately decimating the signals and then quantizing them. To further increase the effective quantizer resolution, noise shaping is introduced by optimizing prefilters and postfilters around the quantizer. We start with a scalar time-invariant quantizer and study two important cases of linear time invariant (LTI) filters, namely, the case where the postfilter is the inverse of the prefilter and the more general case where the postfilter is independent from the prefilter. Closed-form expressions for the optimum filters and average minimum mean square error are derived in each case for both the single band and multiband models. Due to the statistical nature of the signals of interest, the class of noise shaping filters and quantizers is then enlarged to include linear periodically time varying $(LPTV)_M$ filters and periodically time-varying quantizers of period M. Because the general $(LPTV)_M$ case is difficult to track analytically, we study two special cases in great detail and give complete solutions for both the single band and multiband models. Examples are also provided for performance comparisons between the LTI case and the corresponding $(LPTV)_M$ one.**

***Index Terms*—Multirate signal processing, noise shaping, oversampling, PCM techniques, sampling theory.**

I. Introduction

IT IS WELL KNOWN that if a continuous time signal $x(t)$ is σ-bandlimited, then it can be recovered uniquely from its samples $x(nT)$ as long as $T \leq \pi/\sigma$. Extensions of the lowpass sampling theorem such as the bandpass, nonuniform, and derivative sampling theorems can be found in [1]. Recently,

Manuscript received December 19, 1996; revised July 31, 1998. This work was supported in part by the Office of Naval Research under Grant N00014-93-1-0231, Tektronix, Inc., and Rockwell International. The associate editor coordinating the review of this paper and approving it for publication was Dr. Troung Q. Nguyen.

J. Tuqan is with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA.

P. P. Vaidyanathan is with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA.

Publisher Item Identifier S 1053-587X(99)00758-8.

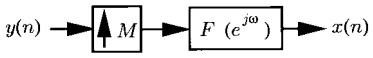


Fig. 1. Single band model.

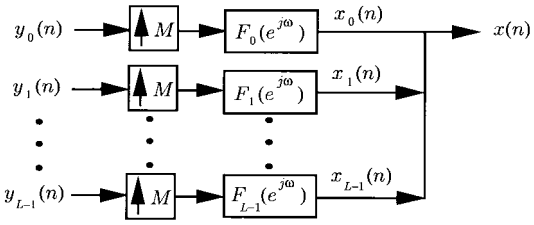


Fig. 2. Multiband model.

Walter [2] showed that under some conditions, a class of nonbandlimited continuous-time signals can be reconstructed from uniformly spaced samples even though frequency aliasing occurs. Vaidyanathan and Phoong [3], [4] developed the discrete-time version of Walter's result from a multirate digital filtering perspective. In specific, they considered the class of nonbandlimited signals that can be modeled as the output of a single *finite order* interpolation filter (single-band model) as in Fig. 1 or as the output of the more general multiband model of Fig. 2. Even though $x(n)$ is not bandlimited (because the interpolation filters are of finite order), it is natural to expect that it can be recovered from its decimated version $x(Mn)$. As a simple example, assume that $x(n)$ is modeled as in Fig. 1. If $F(e^{j\omega})$ is a Nyquist(M) filter (see [5, pp. 151–152]), then $x(Mn)$ is equal to $y(n)$, and we have the relation $x(n) = \Sigma_k x(kM) f(n - kM)$. In other words, $x(n)$ is completely defined by the samples $x(Mn)$, even though the filter $F(e^{j\omega})$ is not ideal, and frequency aliasing occurs. In [4], the authors consider the case where $F(e^{j\omega})$ is not necessarily a Nyquist(M) filter and show how similar reconstruction can be done. They also consider the stability of the reconstruction process.

In this paper, we study the efficient quantization of this class of nonbandlimited signals that can be *accurately* modeled as in Fig. 1 or more generally as in Fig. 2. To motivate such a study, consider the schematic shown in Fig. 3, where the box labeled Q is a simple uniform roundoff (PCM) quantizer. After going through the quantizer, the signal $x(n)$ is now contaminated by an additive noise component $e(n)$. Assuming that the signal $x(n)$ is bandlimited or equivalently oversampled (since a bandlimited signal can be further downsampled), we can lowpass

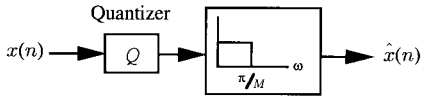


Fig. 3. Schematic of the oversampling PCM technique.

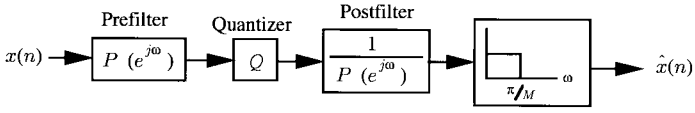


Fig. 4. Quantization scheme of Fig. 3 with noise shapers.

filter the quantized signal $x(n)+e(n)$. The ideal lowpass filter on the right removes the noise in the stopband but does not change the signal component. In terms of signal and noise power, the signal power remains unchanged, whereas the noise power decreases proportionally to the oversampling ratio. It can be shown that for every doubling of the oversampling ratio, the signal-to-noise ratio (SNR) improves by about 3 dB, or equivalently, the quantizer resolution improves by one half bit (see for example [6]). After lowpass filtering, the quantized signal can be downsampled to the Nyquist rate without affecting the SNR. The idea is therefore to exploit the oversampled nature of the signal $x(n)$ to tradeoff quantizer complexity for higher resolution. This technique is usually called oversampled PCM conversion. Consider now the system of Fig. 4, where $P(e^{j\omega})$ is a linear time-invariant (LTI) filter. The input signal $x(n)$ is still assumed to be oversampled (bandlimited). In addition to the benefits described above, it can be shown that this more sophisticated system produces a further decrease in the noise power by "cleverly" choosing the filter $P(e^{j\omega})$ in Fig. 4. The filter pair $P(e^{j\omega})$ and $1/P(e^{j\omega})$ does not modify the input signal $x(n)$ in any way but only affects the noise component $e(n)$. Similar to sigma–delta quantizers, the system of Fig. 4 introduces *noise shaping* in the signal band to allow higher resolution quantization of bandlimited signals.

With these ideas in mind, observe now the output $x(n)$ of Fig. 1. Even though $x(n)$ is not bandlimited, it can be reconstructed from its decimated version as explained above. In this sense, *it can be considered as an oversampled signal*. A question then arises: Can we obtain advantages similar to the above schemes for a nonbandlimited signal satisfying the model of Fig. 1 and, more generally, of Fig. 2? Furthermore, for a fixed set of filters $F(e^{j\omega})$ (or $F_k(e^{j\omega}), k=0,1,\cdots,L-1$), what is the optimum filter $P(e^{j\omega})$ that minimizes the noise power at the output? Do we gain more by using a more general postfilter $V(e^{j\omega})$ instead of $1/P(e^{j\omega})$? This is a sample of the type of questions we answer in this paper. Indeed, we will show that by replacing the ideal lowpass filter with the correct *nonideal* multirate reconstruction system, we can reap the same quantization advantages, as in the bandlimited case. For example, we will show that under the assumption that $|F(e^{j\omega})|^2$ is Nyquist(M) (we will motivate such an assumption later in the paper), the signal $\hat{x}(n)$ in Fig. 5 is equal to $x(n)$ in the absence of the quantizer, and the entire scheme of Fig. 5 behaves similarly to Fig. 3, except that the lowpass filtering is now *multirate* and *nonideal*. Generally

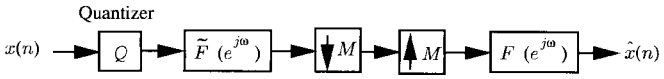


Fig. 5. Multirate quantization scheme for the single-band case.

speaking, if a nonbandlimited signal can be reconstructed from its decimated version $x(Mn)$ because it satisfies a model like Figs. 1 or 2, then a low-precision quantizer should allow us to produce a high-precision version $\hat{x}(n)$.

To bring the analogy closer to the scheme of Fig. 4, we should introduce noise shaping. This can be done by using a prefilter and postfilter before and after the quantizer, respectively, as shown in Fig. 6. The prefilter $P(e^{j\omega})$ is traditionally an integrating lowpass filter. The postfilter $1/P(e^{j\omega})$ shapes the noise spectrum in order to further decrease the noise variance. Several extensions to the above noise shaping idea are also discussed.

The quantization advantage offered by Figs. 5 and 6 can be useful, for example, in the following realistic engineering scenario. Suppose $x(n)$ is generated at a point where we cannot afford very complex signal processing (e.g., in deep space) and needs to be transmitted to a distant place (e.g., earth station). If we have the knowledge that $x(n)$ admits a satisfactory model like Fig. 1, we can compress it using a very simple lowpass filter $P(e^{j\omega})$ with one or two multipliers and then quantize the output before transmission. The postfilter $1/P(e^{j\omega})$ and the expensive multirate filter are at the receiver end, where the complexity is acceptable.

Assume now that the main aim is to obtain a reduction in the bit rate (number of bits per second) rather than accuracy (number of bits per sample). If we are allowed to perform discrete-time filtering (of arbitrary complexity), we will see that the best approach would be as in Fig. 7. In this setup, we first generate the driver signal $y(n)$ and then quantize it. The signal $\hat{x}(n)$, which is equal to $x(n)$ in the absence of quantization, is then generated. The lower rate signal $y(n)$ in Fig. 7 can be regarded as the principal component signal in an orthonormal subband coder. We will see throughout this paper that by choosing this type of quantization system, we can obtain a large reduction in the bit rate and/or the quantization accuracy, depending on the particular signal model.

Summarizing, oversampling PCM conversion and noise shaping are popular techniques that arise in A/D conversion applications but can only be applied to narrowband signals. Indeed, in higher bandwidth applications such as video processing and digital radio, the oversampling requirement has been prohibitive [7]. In [8], the authors propose a parallel architecture wherein multiple sigma–delta modulators are combined so that time oversampling is not required. Instead, the system achieves the effect of oversampling (adding redundant samples) from the multiplicity of the sigma–delta modulators. Our approach in this paper is based on modeling the signal of interest as the output of a single FIR interpolation filter (single-band model) or, more generally, as the sum of the outputs of L FIR interpolation filters (multiband model). The conventional bandlimited scenario described above is then the special case when the filters $F_k(e^{j\omega})$ are ideal filters. The main contribution of the paper is to show how to take advantage of

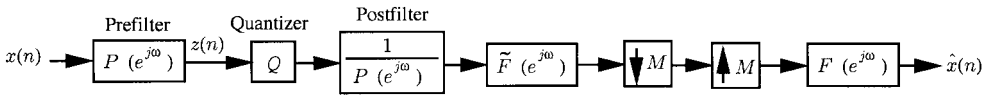


Fig. 6. Noise shaping by LTI prefilters and postfilters for the single-band case where the postfilter is assumed to be the inverse of the prefilter.

these signal models (Figs. 1 and 2) in preparing a quantized or compressed version of $x(n)$. We find that the choice of a particular scheme depends on how much processing we are allowed to do before quantization. If processing is allowed, we first generate $y(n)$ by filtering and decimation and then quantize it. Otherwise, we quantize $x(n)$ directly and then filter the quantized signal with the appropriate multirate scheme. Noise shaping can be also introduced to obtain better resolution. In any case, an improvement in accuracy and/or bit rate due to the signal models is always achieved. The results presented here are therefore a generalization of well-known efficient A/D conversion techniques that apply only to bandlimited signals.

A. Main Results and Outline of the Paper

1) In Section II, definitions and well-established facts of various multirate and statistical signal processing concepts used throughout the paper are reviewed.
2) In Section III, we discuss briefly the multirate modeling of the signal $x(n)$. To be specific, we argue that for an arbitrary input $x(n)$, finding a multiband model is equivalent to the design of a principal component filter bank (PCFB), and finding a single-band model is equivalent to the design of an energy compaction filter.
3) In Section IV, new results that describe the statistical behavior of signals as they pass through multirate interconnections are presented. These results are then used to derive the remaining theorems in the paper.
4) In Section V, we give several results on the quantization of the nonbandlimited signal $x(n)$ modeled as in Fig. 1. The signal $x(n)$ is first quantized to an average of b bits/sample and then filtered by the multirate interconnection in Fig. 5. We show that the multirate system does not affect the signal component but reduces the noise variance by a factor of M. This amounts to the same quantitative advantage obtained from the oversampling PCM technique (0.5 bit reduction per doubling of the oversampling ratio).
5) In Section VI, the lower rate signal $y(n)$ is quantized instead of $x(n)$. By quantizing $y(n)$ to b bits per sample, the quantization bit rate (number of bits per second) is decreased by a factor of M, but noise reduction due to multirate filtering is now not possible.
6) In Section VII, noise shaping is introduced in order to obtain better accuracy. First, we consider the use of pre-and post-linear time-invariant filters $P(e^{j\omega})$ and $1/P(e^{j\omega})$, as in Fig. 6, together with a fixed time-invariant quantizer $\mathcal{Q}$. For this case, the optimum filter $P_{\rm opt}(e^{j\omega})$ that minimizes the quantization noise variance in the reconstructed output $\hat{x}(n)$ is derived, and a closed-form expression for the average minimum mean square error is obtained. We then consider the more general prefilters and postfilters $P(e^{j\omega})$ and $V(e^{j\omega})$, as in Fig. 8. Closed-form expressions for the optimum filters and the average minimum mean square error are also found for this case. We would like to warn the reader at this point that no optimization of finite-order filters is performed in this paper. We derive and use the expressions of the theoretically optimum filters (without order constraint) to get an upper bound on the possible achievable gain.

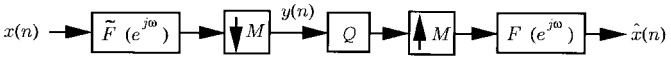


Fig. 7. Quantizing the lower rate signal $y(n)$ (single-band case).

7) In Section VIII, we replace the linear time-invariant filter $P(e^{j\omega})$ with a more general linear periodically time-varying filter of period M. This is motivated by the cyclo-wide-sense stationarity of $x(n)$. Since the problem of finding the optimum general $(LPTV)_M$ filter (equivalently biorthogonal filter bank) is analytically difficult to track, optimal solutions are given for two special cases of $(LPTV)_M$ filters. The first solution is for the set of M filters $V_k(e^{j\omega})$ shown in Fig. 9. The filters $V_k(e^{j\omega})$ and $(1/V_k(e^{j\omega}))$ act as pre- and post-filters for the kth subband quantizer. The second solution is for the case of an orthonormal filter bank, or equivalently, for a lossless $(LPTV)_M$ filter. The scheme is shown in Fig. 10 for the single-band case.
8) All the results mentioned above are also generalized for the multiband case. Furthermore, examples are provided whenever necessary for illustrative purposes.

II. Summary of Standard Multirate Concepts

1) Notations: Lowercase letters are used for scalar time domain sequences. Uppercase letters are used for transform domain expressions. Bold faced quantities represent vectors and matrices. The superscripts $T, *$, and $\dagger$ denote, respectively, the transpose, conjugate, and the conjugate transpose operations for vectors and matrices. The M-fold downsampler has an input–output relation $y(n) = x(n)\downarrow_M = x(Mn)$. The M-fold expander's input–output relation is $y(n) = x(n)\uparrow_M = x(n/M)$ when n is a multiple of M and $y(n) = 0$ otherwise. The M-fold polyphase representation of $X(e^{j\omega})$ is given by $X(e^{j\omega}) = X_0(e^{jM\omega}) + e^{-j\omega}X_1(e^{jM\omega}) + e^{-j2\omega}X_2(e^{jM\omega}) + \cdots + e^{-j(M-1)\omega}X_{M-1}(e^{jM\omega})$. The polyphase components are given by $x_k(n) = x(Mn+k)$ or in the frequency domain by $X_k(e^{j\omega}) = (e^{j\omega k}X(e^{j\omega}))\downarrow_M$. The tilde accent on a function $\boldsymbol{F}(z)$ is defined such that $\tilde{\boldsymbol{F}}(z)$ is the conjugate transpose of $\boldsymbol{F}(z)$, i.e., $\tilde{\boldsymbol{F}}(z) = \boldsymbol{F}^{\dagger}(1/z^*)$.

2) Blocking a Signal: Given a scalar signal $x(n)$, we define its M-fold blocked version $\boldsymbol{x}(n)$ by

$$\boldsymbol{x}(n) = (x(nM) \quad x(nM-1) \quad \cdots \quad x(nM-M+1))^T. \tag{1}$$

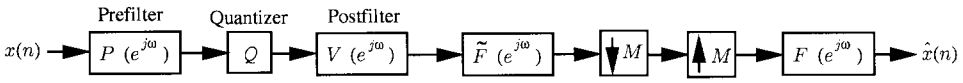


Fig. 8. General LTI prefilters and postfilters for noise shaping for the single-band case.

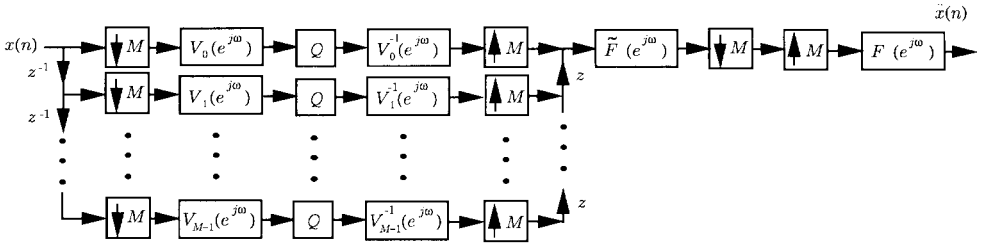


Fig. 9. Scheme 1 for noise shaping using $(LPTV)_M$ prefilters and postfilters (the single-band case).

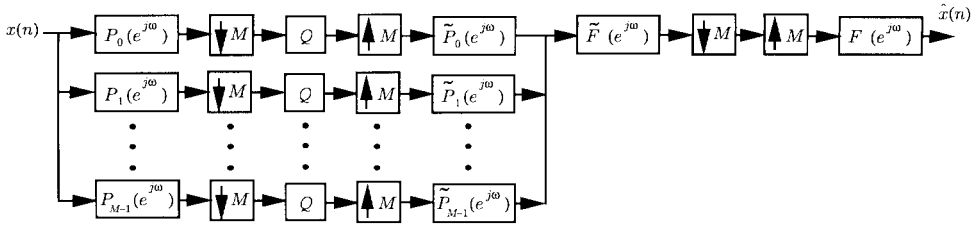


Fig. 10. Scheme 2 for noise shaping using $(LPTV)_M$ prefilters and postfilters (the single-band case).

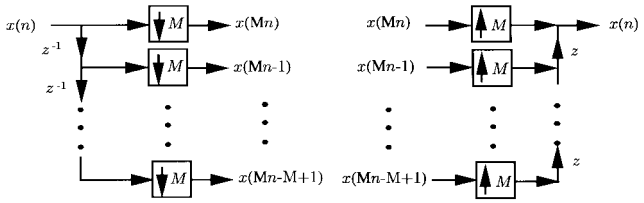


Fig. 11. M-fold blocking of a signal and unblocking of an $M \times 1$ vector signal.

Equivalently, the scalar sequence $x(n)$ is called the unblocked version of the vector process $\boldsymbol{x}(n)$. The blocking and unblocking operations are shown in Fig. 11. The elements of the blocked version $\boldsymbol{x}(n)$ are the polyphase components of $x(n)$.

3) Cyclo-Wide-Sense Stationary Process: A stochastic process $x(n)$ is said to be cyclo-wide-sense stationary with period M ($(CWSS)_M$,) if the M-fold blocked version $\boldsymbol{x}(n)$ is WSS. Alternatively [9], [10], a process $x(n)$ is $(CWSS)_M$ if the mean and autocorrelation functions of $x(n)$ are periodic with period M, i.e.,

$$\begin{aligned} E[x(n)] &= E[x(n+kM)] \quad \forall n,k \quad \text{and} \\ R_{xx}(n,k) &= R_{xx}(n+M,k) \quad \forall n,k \end{aligned} \tag{2}$$

where $R_{xx}(n,k) \triangleq E[x(n)x^*(n-k)]$ is the autocorrelation function of $x(n)$.

4) Antialias(M) Filters: $F(e^{j\omega})$ is said to be an antialias(M) filter if its output can be decimated M-fold without aliasing, no matter what the input is. Equivalently, there is no overlap between the plots $F(e^{j(\omega-(2\pi k/M))})$ for distinct k in $0 \le k \le M-1$. Since this requires a stopband with infinite attenuation, antialias are ideal filters.

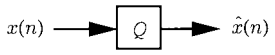


Fig. 12. Direct quantization of $x(n)$.

5) Orthonormal Filter Bank: An M-channel maximally decimated uniform filter bank (FB) is said to have the perfect reconstruction (PR) property when $\boldsymbol{R}(e^{j\omega}) = \boldsymbol{E}^{-1}(e^{j\omega})$, where $\boldsymbol{E}(e^{j\omega})$ and $\boldsymbol{R}(e^{j\omega})$ denote, respectively, the analysis and synthesis polyphase matrices [5]. In the case of an orthonormal filter bank, the analysis polyphase matrix is paraunitary, i.e., $\boldsymbol{E}(e^{j\omega})\boldsymbol{E}^{\dagger}(e^{j\omega}) = \boldsymbol{I}\ \forall\omega$, and we choose $\boldsymbol{R}(e^{j\omega}) = \boldsymbol{E}^{\dagger}(e^{j\omega})$ for perfect reconstruction. The analysis and synthesis filters are related by $F_k(e^{j\omega}) = \tilde{H}_k(e^{j\omega})$, that is $f_k(n) = h_k^*(-n)$. It follows that for an orthonormal filter bank, the energy of each analysis/synthesis filter equals unity, that is, $\int_{-\pi}^{\pi} |F_k(e^{j\omega})|^2 d\omega/2\pi = 1$.

6) The Coding Gain of a System: Assume that we quantize $x(n)$ directly with b bits, as shown in Fig. 12. We denote the corresponding mean square error (mse) by $\mathcal{E}_{\text{direct}}$. We then use the optimum pre- and post-filters (in the mean square sense) around the quantizer. With the rate of the quantizer fixed to the same value b, we denote the minimum mse in this case by $\mathcal{E}_{\min}$. The ratio $\mathcal{E}_{\text{direct}}/\mathcal{E}_{\min}$ is called the coding gain of the new system and, as the name suggests, is a measure of the benefits provided by the pre/postfiltering operation.

III. Multirate Signal Modeling

In this paper, we are interested in the multirate modeling of a WSS random process, say $x(n)$, as in Fig. 1 or, more generally, as in Fig. 2. The signal $y(n)$ in Fig. 1 is a zero mean WSS process, and the signals $y_k(n), k = 0, 1, \cdots, L-1$,

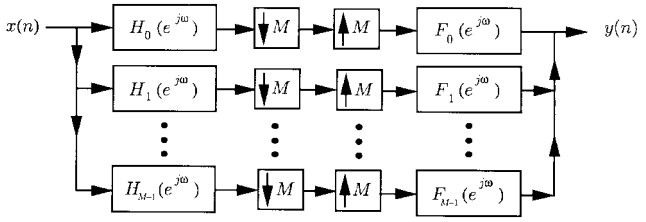


Fig. 13. M-channel FIR maximally decimated uniform filter bank.

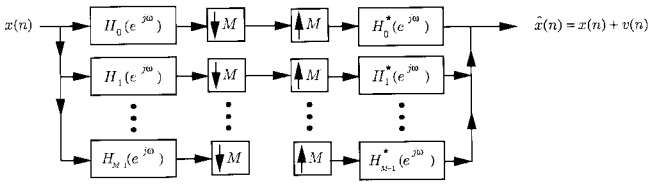


Fig. 14. M-channel FIR principal component filter bank with $P = 2$.

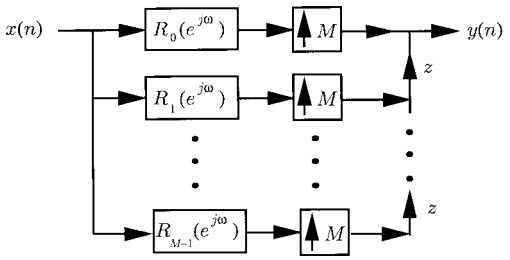


Fig. 15. Equivalent polyphase representation of Fig. 1.

in Fig. 2 are assumed to be zero mean jointly WSS random processes. In both cases, the model filter(s) $F_k(e^{j\omega})$ are assumed to be FIR. Note that the output $x(n)$ is, in general, a zero mean cyclo-wide-sense stationary random process of period $M(CWSS)_M$ [10]. In fact, $x(n)$ is WSS if, and only if, the model filters $F_k(e^{j\omega})$ are antialias(M) filters. Therefore, unlike in standard stochastic rational modeling (e.g., AR, MA, and ARMA modeling), a WSS signal in this case is "approximated" by a $(CWSS)_M$ signal.

A. Finding a Signal Model

What kind of signals can be realistically modeled as in Fig. 1 or, more generally, as in Fig. 2? To answer this, consider the filter bank system of Fig. 13, where a WSS signal $x(n)$ is split into M subbands and reconstructed perfectly from its maximally decimated versions. Suppose now that the signal $x(n)$ has most of its energy concentrated in L subbands, which we number as the first L subbands. Then, the signal model of Fig. 2 is a good approximation of the original signal. Similarly, if the signal $x(n)$ has a lowpass or bandpass spectrum with most of its energy concentrated in a bandwidth of $2\pi/M$, then we can accurately represent the original signal $x(n)$ with the signal model of Fig. 1. Thus, given a signal $x(n)$ with energy concentrated mostly in certain subbands, the problem of finding the best signal model reduces to that of finding the filter bank that produces the L most dominant subbands. If the filter bank in Fig. 13 is orthonormal (paraunitary), the modeling issue reduces to the design of the so-called principal component filter banks for the multiband case and the design of energy compaction filters for the single band case. These important concepts are discussed next.

B. Principal Component FB's and Energy Compaction Filters

Consider Fig. 14, where $(M-P)$ channels are dropped in the synthesis part of an M-channel *orthonormal* filter bank. An orthonormal filter bank that minimizes the average mean square reconstruction error for *all* P is called a principal component filter bank (PCFB) [11]. By definition, it can be shown that a PCFB produces a decreasing arrangement of the the subband variances $\sigma_{y_0}^2 \geq \sigma_{y_1}^2 \cdots \geq \sigma_{y_P}^2$ such that for all $1 \leq P < M, \Sigma_{k=0}^{P-1} \sigma_{y_k}^2$ is maximized. For $P = M, \Sigma_{k=0}^{M-1} \sigma_{y_k}^2 = M\sigma_x^2$ and is therefore fixed. The set of subband variances $\{\sigma_{y_k}^2\}$ generated by a PCFB is said to "majorize" any other arbitrary set of subband variance $\{\sigma_{x_k}^2\}$. For the case of $P = 1$, the problem becomes one of designing a single analysis filter such that its output variance is maximized under the constraint that its magnitude squared response is Nyquist(M). The resulting filter is termed an energy compaction filter.

A procedure that finds the *globally* optimal FIR energy compaction filter $F(e^{j\omega})$ for any M and arbitrary filter order N can be found in [12] and [13]. Depending on the FIR filter order N, an input $x(n)$ can be very accurately modeled as in Fig. 1. The tradeoff between the original signal $x(n)$ and its model representation becomes one of accuracy versus efficiency, which is typical in signal modeling applications. The design of *globally* optimal FIR principal component filter bank remains at this moment in time an open problem (see [14] for some preliminary results). PCFB's and energy compaction filters play a key role in the optimization of an orthonormal filter bank according to the input second-order statistics. The above ideas therefore find applications in the area of subband coding, i.e., the optimization of orthonormal filter banks taking into account the effect of subband quantization. A full description of the various connections between PCFB's, energy compaction filters, and the subband coding problem is beyond the scope of this paper; we refer the reader to [12] for more details on this subject.

C. Filter and Quantizer Assumptions

Filter Assumptions: Based on the previous discussion, the finite-order filter $F(e^{j\omega})$ of Fig. 1 is assumed to be an optimum energy compaction filter, and $y(n)$ is the subband signal corresponding to the most dominant subband. Similarly, the finite-order filters $F_k(e^{j\omega}), k = 0, 1, \cdots, L-1$ of Fig. 2 are assumed to be the L-first synthesis filters of a principal component filter bank, and $y_k(n)$ are the subband signals corresponding to the L most dominant subbands. Although this particular choice minimizes the approximation (modeling) error, we emphasize that this choice is not *necessary* for developing the results of this paper.

Quantizer Assumption: As a convention for this paper, the box labeled $\mathcal{Q}$ represents a scalar uniform (PCM) quantizer and is modeled as an additive zero mean white noise source $q(n)$. Because the model filters are not ideal, the input $x(n)$ is a zero

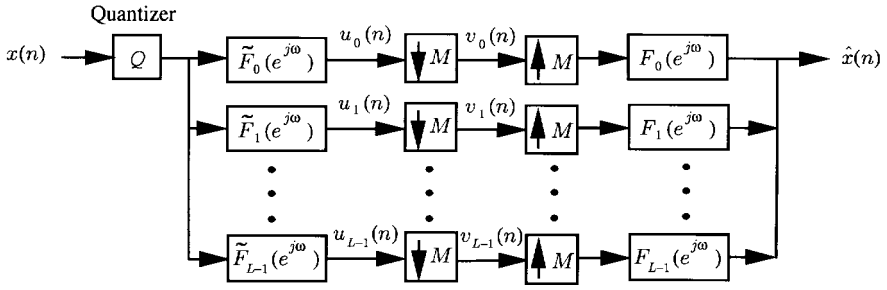


Fig. 16. Multirate quantization scheme for the multiband model.

mean $(CWSS)_M$ process. Since the input to the quantizer $x(n)$ is a $(CWSS)_M$ process, its variance $\sigma_x^2(n)$ is a periodic function of n with period M. Define σ_x^2 to be the average variance of $x(n)$, i.e., $\sigma_x^2 = (1/M)\Sigma_{n=0}^{M-1}\sigma_x^2(n)$. Then, choose the fixed step size Δ in the uniform quantizer such that the quantization noise variance σ_q^2 is directly proportional to the average variance of the quantizer input $x(n)$, that is

$$\sigma_q^2 = c2^{-2b}\sigma_x^2 \tag{3}$$

where

- σ_q^2 quantization noise variance;
- c constant that depends on the statistical distribution of $x(n)$ and the overflow probability;
- σ_x^2 average variance of the quantizer input.

The above relation is justified for a PCM quantizer using three (or more) bits per sample (see [15, ch. 4]). If the input to $\mathcal{Q}$ is WSS, the above relation holds with σ_x^2 now denoting the actual variance of the WSS process.

IV. Preliminary Results

Result 1: Consider any L synthesis filters $(L<M)$ of an M-channel orthonormal filter bank as shown in Fig. 2. Assume that the L inputs $y_k(n)$ to the synthesis filters $F_k(e^{j\omega})$ are zero mean jointly WSS processes that are not necessarily uncorrelated. Then, the statistical correlation (averaged over M samples) between the interpolated subband signal $x_i(n)$ and the M-sample shifted process $x_j(n-Mm)$ is zero for all values of $i \neq j$ and m, that is

$$\frac{1}{M}\sum_{k=0}^{M-1} E[x_i(n-k)x_j^*(n-k-Mm)] = 0, \forall n, m \quad \text{and} \quad \forall i, j \in [0, L-1], i \neq j. \tag{4}$$

The proof can be found in Appendix A. As a consequence, the average variance of the $(CWSS)_M$ output process $x(n)$ of Fig. 2, where the filters $F_k(e^{j\omega})$ are any L synthesis filters of an M-channel orthonormal filter bank, is

$$\sigma_x^2 = \frac{1}{M}\sum_{k=0}^{L-1}\sigma_{y_k}^2. \tag{5}$$

This can be seen by substituting $x(n)$ in the formula $\sigma_x^2 = (1/M)\,\Sigma_{n=0}^{M-1}\,E[|x(n)|^2]$ and using result 1 for the special case of $m=0$ and $n=M-1$. If the L inputs to the synthesis filters $F_k(e^{j\omega})$ are zero mean uncorrelated WSS processes, the previous result holds without the orthonormality requirement on the filters $F_k(e^{j\omega}), k = 0, 1, \cdots, L-1$.

Result 2: Consider the multirate interconnection of Fig. 1, where the input $y(n)$ is zero mean WSS random process. If $F(e^{j\omega})$ is a filter (not necessarily ideal) with a Nyquist(M) magnitude squared response, then

$$\sigma_x^2 = \frac{1}{M}\sigma_y^2 \tag{6}$$

where σ_x^2 is the average variance of the $(CWSS)_M$ output $x(n)$.

Proof: While this is a special case of the above with $L=1$, the following proof is direct and more instructive. With $F(e^{j\omega})$ expressed in terms of its polyphase components $R_k(e^{j\omega})$, Fig. 1 can be redrawn as in Fig. 15. The signal $x(n)$ is the interleaved version of the WSS outputs of $R_k(e^{j\omega})$. Therefore, it has zero mean and a variance that is periodic with period M. The average variance is given by

$$\sigma_x^2 = \frac{1}{M}\sum_{k=0}^{M-1}\sigma_{x_k}^2 = \frac{1}{M}\int_{-\pi}^{\pi} S_{yy}(e^{j\omega})\sum_{k=0}^{M-1}|R_k(e^{j\omega})|^2\,\frac{d\omega}{2\pi}. \tag{7}$$

The Nyquist property of $|F(e^{j\omega})|^2$ implies, in particular, that $\Sigma_{k=0}^{M-1}|R_k(e^{j\omega})|^2 = 1$ (see [5, p. 159]). The preceding equation therefore simplifies to $\sigma_x^2 = (1/M)\int_{-\pi}^{\pi} S_{yy}(e^{j\omega})(d\omega/2\pi) = (1/M)\sigma_y^2$. ■

V. Increasing the Quantizer Resolution by Multirate Filtering

Consider the set up shown in Fig. 5 for the single-band model and in Fig. 16 for the multiband case. In the absence of the quantization, the two schemes are PR systems. In the presence of the quantizer, the output $\hat{x}(n)$ in Figs. 5 and 16 is equal to the original sequence $x(n)$ plus an error signal $e(n)$ due to quantization. The following result shows that by using the above schemes, a significant reduction in the average mean square error $\mathcal{E} \triangleq (1/M)\,\Sigma_{n=0}^{M-1}\,E\{e(n)\}^2$ can be obtained in comparison with the direct quantization of $x(n)$ shown in Fig. 12.

Theorem 5.1: Consider the scheme of Fig. 16, where the L filters $F_k(e^{j\omega})$ are assumed to be any L channels of an M-channel critically sampled orthonormal FB. Under the above quantization noise assumption, the average mse $\mathcal{E} \triangleq (1/M)\,\Sigma_{n=0}^{M-1}\,E\{\hat{x}(n)-x(n)\}^2$ is equal to $(L/M)\sigma_q^2$.

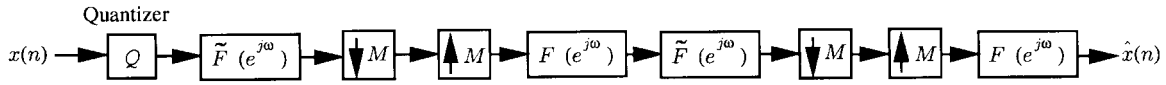


Fig. 17. Cascade of two multirate interconnections for the single band case.

Proof: Because the system is a PR one in the absence of quantization, the average error at the output is due only to the quantization noise. The quantization noise $q(n)$ is white and propagates through the L channels of Fig. 16. For the kth channel, the variance of $u_k(n)$ due to the noise passage through $F_k(e^{j\omega})$ is given by

$$\sigma_{u_k}^2 = \sigma_q^2 \int_{-\pi}^{\pi} |F_k(e^{j\omega})|^2 \frac{d\omega}{2\pi} = \sigma_q^2. \tag{8}$$

The second equality follows because the filters have unit energy. The downsampling operation does not alter the variance of a signal. We therefore obtain $\sigma_{v_k}^2 = \sigma_{u_k}^2 = \sigma_q^2$ for all k. Using result 1 of Section III, we can write

$$\mathcal{E} = \frac{1}{M} \sum_{k=0}^{L-1} \sigma_{v_k}^2 = \frac{L}{M} \sigma_q^2. \tag{9}$$

■

For the scheme of Fig. 5, the average mse $\mathcal{E}$ can be obtained directly by setting $L = 1$ and is therefore equal to $(1/M)\sigma_q^2$. The quantization noise variance σ_q^2 obtained by directly quantizing $x(n)$, as shown in Fig. 12, is now reduced by the oversampling factor M. The signal variance σ_x^2, on the other hand, did not change. By expressing the interpolator M in the form 2^r, we can immediately see that we can get the same quantitative advantage of the oversampling PCM technique, namely, an increase in SNR by 3 dB for every doubling of the oversampling factor. For example, for the single-band case of Fig. 5, if $M = 2$, then we get an SNR increase of 3 dB, whereas if $M = 4$, the SNR increment is by 6 dB. Some important remarks are in order at this point:

1) In the oversampling PCM technique, the quantized bandlimited signal is typically downsampled after the lowpass filter [6]. The SNR before and after the downsampler is the same, and the increase in SNR is only due to a reduction in noise power. Similarly, the SNR before and after the interpolation filter in Fig. 5 does not change. However, the reason for the SNR increase before the interpolation filter is different from the one after the interpolation filter. To be specific, at the input of the interpolation filter, the signal variance increases proportionally to M since $\sigma_y^2 = M\sigma_x^2$, and the noise power remains fixed. At the output of the interpolation filter, the signal variance does not change, but the noise power decreases in proportion to M. In both cases, this amounts to the same SNR improvement. This last technical difference arises because our study assumes a statistical framework rather than a deterministic one (typical in A/D conversion applications) and because of our quantizer assumptions.
2) *Intuitive Explanation of Theorem 5.1*: The signal $x(n)$, which is modeled either as in Figs. 1 or 2, is oversampled and, therefore, contains redundant information in the form of an excess of samples. It is by quantizing these extra samples that we obtain the reduction in the quantization noise variance (equivalently in the average mean square error). We are therefore effectively quantizing with a higher number of bits per sample. This tradeoff between the quantization noise variance (effective quantizer resolution) and the sampling rate is the underlying principle of oversampled A/D converters.
3) *The Role of the Factor L in This Analysis*: The parameter L, which is defined to be the number of channels in the multiband case, alternates between two extremes: $L = 1$ and $L = M$. When $L = 1$, we get the best SNR improvement at the expense of a more narrow class of inputs $x(n)$. When $L = M$, it is clear from (9) that no noise variance reduction is achieved since the class of signals is now unrestricted. We can also see this by noticing that the multirate interconnection in Fig. 16 becomes a PR filter bank that is signal independent. The parameter L therefore determines the tradeoff between the generality of the class of signals $x(n)$ and the reduction in quantization noise variance.
4) *A Cascade of the Scheme of Fig. 5 Does Not Provide Any Further Gain*: Using the scheme of Fig. 5, we obtained a reduction in noise by a factor M. If we use a cascade of the same filtering scheme as in Fig. 17, no further noise reduction is obtainable. Using the polyphase identity [5] and keeping in mind that $|F(e^{j\omega})|^2$ is Nyquist(M), the product filter $F(e^{j\omega})\tilde{F}(e^{j\omega})$ together with the expander and decimator reduces to an identity system. Fig. 17 therefore simplifies to Fig. 5, and the average mse is the same.
5) *Interpretation on Terms of Projection Operators*: The last comment (Remark 4) indicates that the filtering scheme in Fig. 5 is a *projection* operator. Therefore, the reduction in noise variance can be attributed to the following line of reasoning: Assume that the filter $F(e^{j\omega})$ corresponds to one of the subband filters in an M-channel orthonormal filter bank. Then, the noise variance $q(n)$ has the following orthonormal expansion: $q(n) = \Sigma_{i=0}^{M-1} \Sigma_k q_i(k) f_i(n - kM)$. The noise signal $e(n)$ at the output of Fig. 5 is obtained by discarding $M - 1$ signals $q_i(k)$ and is therefore an orthogonal projection of $q(n)$ onto the subspace spanned by the filter $F(e^{j\omega})$ only.

VI. Quantizing at Lower Rate

A consequence of the previous results and discussion is then the natural question: What if the discrete time filtering of the oversampled signal is not a major burden? If we know that $x(n)$ can be modeled quite accurately by the filter $F(e^{j\omega})$ of Fig. 1 or the filters $F_k(e^{j\omega}), k = 0, 1, \cdots, L-1$ of Fig. 2, we can filter and downsample $x(n)$ accordingly to obtain either $y(n)$ or $y_k(n), k = 0, 1, \cdots, L-1$. The

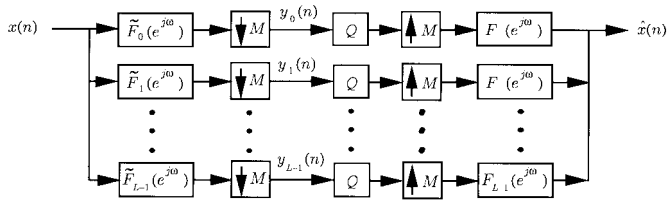


Fig. 18. Quantizing the lower rate signals $y_k(n)$ (multiband case).

quantization systems for the two models are shown in Figs. 7 and 18, respectively. We can then in principle quantize the decimated signal $y(n)$ in Fig. 7 with $\hat{b} = Mb$ bits/sample or the signals $y_k(n), k = 0, 1, \cdots, L-1$ of Fig. 18 with an average number of bits per sample $\hat{b} = (M/L)b$ bits. This situation is equivalent to fixing the bit rate (number of bits per second) to be equal to b in order to trade quantization resolution with sampling rate. Moreover, for the multiband case, we can allocate bits b_k to the driving signals $y_k(n)$ in an "appropriate" manner. At this point, we will, however, assume that the goal is to actually obtain a reduction in the bit rate. To achieve this, we let $\hat{b}$ be equal to b for both cases and analyze the quantization systems of Figs. 7 and 18 under this condition. By fixing the number of bits per sample and decreasing the signal rate, the bit rate will automatically decrease by M/L. However, since the quantizer resolution did not increase, the quantization noise variance should not differ from the direct quantization case of Fig. 12. This last statement is verified formally in the next theorems.

Theorem 6.1: Consider the scheme of Fig. 7. Using a fixed number of bits per sample b to quantize $y(n)$, the average mean square error $\mathcal{E}$ is equal to σ_q^2, where σ_q^2 is the noise variance obtained from directly quantizing $x(n)$ using b bits per sample.

Proof: Let σ_q^2 be the noise variance of Fig. 12 and $\mathcal{E}$ be the average mean square error of Fig. 7. Using (3), we can write $\sigma_q^2 = c2^{-2b}\sigma_x^2$. However, by Result 2 of Section III, $\mathcal{E} = (1/M)c2^{-2b}\sigma_y^2 = (1/M)c2^{-2b}M\sigma_x^2 = \sigma_q^2$, where σ_x^2 is the average variance of $x(n)$. ■

The theorem indicates that for the single-band model and under a fixed number of quantizer bits b, quantizing the lower rate signal $y(n)$ is as accurate as directly quantizing $x(n)$. This is expected and is, in fact, consistent with the observation of Section V regarding the tradeoff between the average mse due to quantization and the rate of the signal. The next theorem for the multiband case gives a similar conclusion.

Theorem 6.2: Consider the scheme of Fig. 18. Assume that we quantize $y_k(n)$ at b bits/sample for all k. Then, the average mse $\mathcal{E}$ is equal to σ_q^2, where σ_q^2 is the noise variance obtained from directly quantizing $x(n)$ using b bits/sample.

Proof: The average mean square error at the output of Fig. 18 is equal to

$$\mathcal{E} = \frac{1}{M}\sum_{k=0}^{L-1}\sigma_{q_k}^2 = \frac{1}{M}c2^{-2b}\sum_{k=0}^{L-1}\sigma_{y_k}^2 \tag{10}$$

where b denotes the *fixed* number of bits allocated to the kth-channel quantizer. The noise variance σ_q^2 in Fig. 12 is equal to $c2^{-2b}\sigma_x^2$, which, in turn, is equal to (10). ■

VII. Noise Shaping by Time-Invariant Prefilters and Postfilters

Following the philosophy of sigma–delta modulators, we now perform noise shaping to achieve a further reduction in the average mean square error. To accomplish this, we propose using LTI pre- and post-filters around the PCM quantizer, as shown in Fig. 6, for the single-band model and in Fig. 19 for the multiband model. We first use a prefilter $P(e^{j\omega})$ and assume that the postfilter is its inverse. We then relax this condition and assume a more general postfilter $V(e^{j\omega})$. The goal is to optimize these filters such that the average mse at the output of either quantization system is minimized. The noise shaping filters to be optimized are not constrained to be rational functions (i.e., of finite order), and noncausal solutions, for example, are accepted.

Although our quantizer design assumptions are the same as before, the quantizer input is no longer the $(CWSS)_M$ process $x(n)$, but a filtered version of it, which we denote by $z(n)$. Following (3), the noise variance in this case is given by $\sigma_q^2 = c2^{-2b}\sigma_z^2$, where σ_z^2 is the average variance of the process $z(n)$. We emphasize that $z(n)$ is a $(CWSS)_M$ process since the output of a linear time invariant filter driven by a $(CWSS)_M$ process is also $(CWSS)_M$ [10]. It is then possible to express σ_z^2 in terms of the prefilter $P(e^{j\omega})$ and the so-called average power spectral density (see below) of the process $x(n)$, denoted by $\hat{S}_{xx}(e^{j\omega})$, as

$$\sigma_z^2 = \int_{-\pi}^{\pi}|P(e^{j\omega})|^2\hat{S}_{xx}(e^{j\omega})\frac{d\omega}{2\pi}. \tag{11}$$

The proof of (11) can be found in Appendix C. The average power spectral density is a familiar concept that arises when "stationarizing" a $(CWSS)_M$ process [16]–[18] and satisfies the well-known properties of the power spectrum of a WSS process. It is defined to be the discrete-time fourier transform of the time averaged autocorrelation function $\hat{R}_{xx}(k)$ given by $(1/M)\sum_{n=0}^{M-1}E[x(n)x^*(n-k)]$. Another interpretation of the average power spectral density that can be physically more appealing is based on the concept of phase randomization and is reviewed in Appendix B. Finally, if $x(n)$ is modeled as in Fig. 1, it can be shown that

$$\hat{S}_{xx}(e^{j\omega}) = \frac{1}{M}S_{yy}(e^{j\omega M})|F(e^{j\omega})|^2 \tag{12}$$

whereas if the signal satisfies the multiband model of Fig. 2, the average power spectral density takes the form

$$\hat{S}_{xx}(e^{j\omega}) = \frac{1}{M}\boldsymbol{F}^{\dagger}(e^{j\omega})\boldsymbol{S_y}(e^{j\omega M})\boldsymbol{F}(e^{j\omega}) \tag{13}$$

where $\boldsymbol{F}(e^{j\omega}) = (F_0(e^{j\omega}) \;\; F_1(e^{j\omega}) \;\; \cdots \;\; F_{L-1}(e^{j\omega}))^T$, and $\boldsymbol{S_y}(e^{j\omega})$ is the $L\times L$ power spectral density matrix of the L WSS inputs $y_k(n)$. Note that when the signals $y_k(n)$ are uncorrelated, (13) simplifies to $(1/M)\sum_{k=0}^{L-1}S_{y_k}(e^{j\omega M})|F_k(e^{j\omega})|^2$. The proofs of (12) and (13) are given in Appendix D. The expression (12) was derived previously in [10] for the special case where $F(e^{j\omega})$ is an anti-alias(M) filter. Furthermore, the

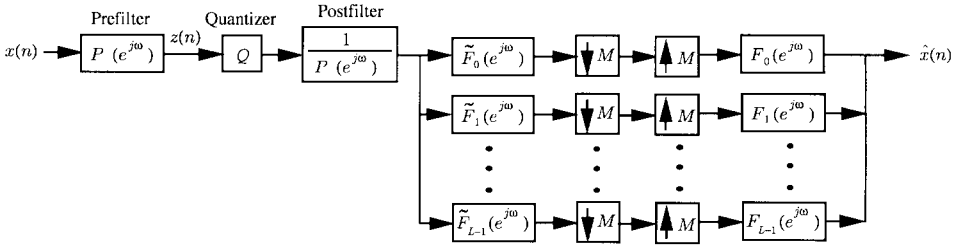


Fig. 19. Noise shaping by LTI prefilters and postfilters for the multiband case where the postfilter is assumed to be the inverse of the prefilter.

authors prove that the output process $x(n)$ is WSS if and only if $F(e^{j\omega})$ is an anti-alias(M) filter. In summary, the statistical properties of the output $x(n)$ of Fig. 1 depend on $F(e^{j\omega})$. If the filter is an anti-alias(M) filter, then $x(n)$ is WSS with a power spectral density $S_{xx}(e^{j\omega})$ in the same form as (12). Otherwise, $x(n)$ is a $(CWSS)_M$ process, and in this case, the average power spectral density $\hat{S}_{xx}(e^{j\omega})$ is given by (12).

A. Case Where the Postfilter is the Inverse of the Prefilter

Theorem 7.1.1: Consider the scheme of Fig. 19 under the same assumptions of Section IV. The optimum prefilter $P(e^{j\omega})$ that minimizes the average mean square reconstruction error has the following magnitude squared response:

$$|P_{\text{opt}}(e^{j\omega})|^2 = \frac{\sqrt{\left(\sum_{i=0}^{L-1}|F_i(e^{j\omega})|^2\right)}}{\sqrt{\hat{S}_{xx}(e^{j\omega})}}. \tag{14}$$

Proof: We first observe that in the absence of quantization, the system of Fig. 19 is a PR system. Therefore, the average mean square reconstruction error $\sigma_e^2 = (1/M)\,\Sigma_{n=0}^{M-1} E[\hat{x}(n) - x(n)]^2$ at the output is due only to the noise signal. Let $v_k(n)$ be the filtered noise component in the kth channel of the L-channel filter bank of Fig. 19. The variance of this signal $\sigma_{v_k}^2$ is equal to

$$\sigma_{v_k}^2 = \int_{-\pi}^{\pi} \sigma_q^2 \frac{|F_k(e^{j\omega})|^2}{|P(e^{j\omega})|^2}\frac{d\omega}{2\pi}. \tag{15}$$

Since the downsampling operation does not change the variance of a process, we can write

$$\sigma_e^2 = \frac{1}{M}\sum_{k=0}^{L-1}\sigma_{v_k}^2 = \sigma_q^2\frac{1}{M}\int_{-\pi}^{\pi}\frac{\sum_{k=0}^{L-1}|F_k(e^{j\omega})|^2}{|P(e^{j\omega})|^2}\frac{d\omega}{2\pi}. \tag{16}$$

Using (3) and (11), we get

$$\sigma_e^2 = \frac{c2^{-2b}}{M}\int_{-\pi}^{\pi}\hat{S}_{xx}(e^{j\omega})|P(e^{j\omega})|^2\frac{d\omega}{2\pi} \cdot \int_{-\pi}^{\pi}\frac{\sum_{k=0}^{L-1}|F_k(e^{j\omega})|^2}{|P(e^{j\omega})|^2}\frac{d\omega}{2\pi}. \tag{17}$$

To find the optimum prefilter $P(e^{j\omega})$, we apply the Cauchy–Schwartz inequality to (17) to obtain

$$\sigma_e^2 \geq \frac{c2^{-2b}}{M}\left(\int_{-\pi}^{\pi}\sqrt{\hat{S}_{xx}(e^{j\omega})\left(\sum_{i=0}^{L-1}|F_i(e^{j\omega})|^2\right)}\frac{d\omega}{2\pi}\right)^2. \tag{18}$$

Since this lower bound is independent of $P(e^{j\omega})$, it is indeed the required minimum and is achieved iff

$$\sqrt{\hat{S}_{xx}(e^{j\omega})}|P(e^{j\omega})| = \frac{\sqrt{\left(\sum_{i=0}^{L-1}|F_i(e^{j\omega})|^2\right)}}{|P(e^{j\omega})|} \tag{19}$$

which gives (14). ■

A number of observations should be made at this point. First, the optimum filter is not unique since the phase response is not specified. Second, the above derivation assumes that the input average spectrum $\hat{S}_{xx}(e^{j\omega}) \neq 0$ for all ω. The assumption is a reasonable one because $x(n)$ is assumed to be nonbandlimited, and therefore, $\hat{S}_{xx}(e^{j\omega})$ cannot be identically zero on a segment of $[0, 2\pi)$. If $\hat{S}_{xx}(e^{j\omega})$ has an isolated zero for some ω, then the resulting prefilter will have a zero on the unit circle and is therefore unstable. In any case, a practical system would use only a stable rational approximation of the ideal solution. Finally, we note that the optimum filter for the scheme of Fig. 6 can be obtained again as a special case by setting $L = 1$ in (14). The optimum prefilter will then have the following magnitude squared response:

$$|P_{\text{opt}}(e^{j\omega})|^2 = \frac{|F(e^{j\omega})|}{\sqrt{\hat{S}_{xx}(e^{j\omega})}} \tag{20}$$

and can be regarded as a multirate extension of the half whitening filter [15]. Using (20), we can derive an interesting expression for the coding gain of the scheme of Fig. 6.

Theorem 7.1.2: With the optimum choice of the prefilter and postfilter, the coding gain expression for the scheme of Fig. 6 is

$$\mathcal{G}_{\text{opt}} = \frac{M\int_{-\pi}^{\pi}S_{yy}(e^{j\omega})\frac{d\omega}{2\pi}}{\left(\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}\frac{d\omega}{2\pi}\right)^2} = M\cdot\mathcal{G}_{hw} \tag{21}$$

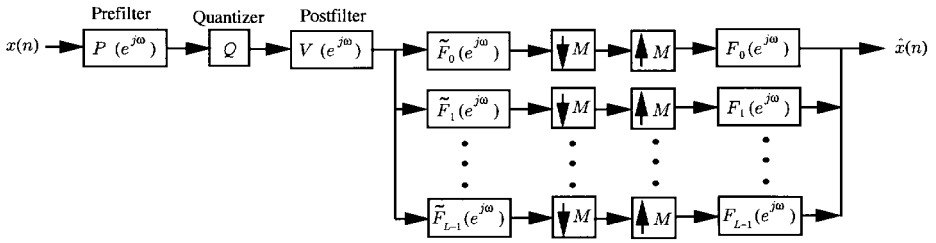


Fig. 20. General LTI prefilters and postfilters for noise shaping for the multiband case.

where $\mathcal{G}_{hw}$ is the half whitening coding gain of the WSS process $y(n)$ [15].

Proof: By definition, the coding gain of the system is given by

$$\mathcal{G}_{\text{opt}} = \frac{\sigma_q^2}{\mathcal{E}_{\text{opt}}} = \frac{\sigma_x^2}{\frac{1}{M}\left(\int_{-\pi}^{\pi}\sqrt{\hat{S}_{xx}(e^{j\omega})}|F(e^{j\omega})|\frac{d\omega}{2\pi}\right)^2} = \frac{M\int_{-\pi}^{\pi}\hat{S}_{xx}(e^{j\omega})\frac{d\omega}{2\pi}}{\left(\int_{-\pi}^{\pi}\sqrt{\hat{S}_{xx}(e^{j\omega})}|F(e^{j\omega})|\frac{d\omega}{2\pi}\right)^2}. \tag{22}$$

Substituting (12) in (22) and simplifying, we get

$$\mathcal{G}_{\text{opt}} = \frac{M\int_{-\pi}^{\pi}S_{yy}(e^{jM\omega})|F(e^{j\omega})|^2\frac{d\omega}{2\pi}}{\left(\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{jM\omega})}|F(e^{j\omega})|^2\frac{d\omega}{2\pi}\right)^2}. \tag{23}$$

The integrals in both the numerator and the denominator can be interpreted as the variance of a WSS random process with a power spectrum density equal to $S_{yy}(e^{jM\omega})|F(e^{j\omega})|^2$ and $\sqrt{S_{yy}(e^{jM\omega})}|F(e^{j\omega})|^2$, respectively. However, we know that downsampling a WSS process produces another WSS process with the same variance. Therefore, we can write

$$\mathcal{G}_{\text{opt}} = \frac{M\int_{-\pi}^{\pi}(S_{yy}(e^{jM\omega})|F(e^{j\omega})|^2)\downarrow_M\frac{d\omega}{2\pi}}{\left(\int_{-\pi}^{\pi}(\sqrt{S_{yy}(e^{jM\omega})}|F(e^{j\omega})|^2)\downarrow_M\frac{d\omega}{2\pi}\right)^2}. \tag{24}$$

Using the fact that $(S_{yy}(e^{jM\omega})|F(e^{j\omega})|^2)\downarrow_M = S_{yy}(e^{j\omega})(|F(e^{j\omega})|^2)\downarrow_M$ and that $(|F(e^{j\omega})|^2)\downarrow_M = 1$, we get (21). ■

The factor M in (21) is again due to the oversampled nature of the signal $x(n)$. It is interesting to note that the noise shaping contribution to $\mathcal{G}_{\text{opt}}$ in (21), which we denote by $\mathcal{G}_{hw}$, is exactly the *coding gain we would obtain by half whitening the WSS process $y(n)$ in the usual way* [15]. By appealing to the Cauchy–Schwartz inequality again, we can show that $\mathcal{G}_{hw} \geq 1$ with equality iff the power spectral density $S_{yy}(e^{j\omega})$ is a constant, i.e., $y(n)$ is white noise. Therefore, for the particular system of Fig. 6, we will not get additional coding gain by noise shaping if the driving WSS process $y(n)$ in Fig. 1 is white noise. For completeness, we would like to mention that

$$\mathcal{G}_{\text{opt}} = \frac{M\int_{-\pi}^{\pi}\sum_{k=0}^{L-1}S_{y_k}(e^{jM\omega})|F_k(e^{j\omega})|^2\frac{d\omega}{2\pi}}{\left(\int_{-\pi}^{\pi}\sqrt{\sum_{i=0}^{L-1}S_{y_i}(e^{jM\omega})|F_i(e^{j\omega})|^2}\sqrt{\sum_{n=0}^{L-1}|F_n(e^{j\omega})|^2}\frac{d\omega}{2\pi}\right)^2} \tag{25}$$

for the coding gain of Fig. 19 (the multiband case) can be derived under the assumption that the JWSS processes $y_k(n), k = 0, 1, \cdots, L-1$ are uncorrelated.

B. Using a More General Postfilter

Consider now the more general system of Fig. 8, where the postfilter is not assumed to be the inverse of the prefilter. The multiband case is shown in Fig. 20. The goal is to jointly optimize the prefilter $P(e^{j\omega})$ and the postfilter $V(e^{j\omega})$ to again minimize the average mse $\triangleq 1/M\,\Sigma_{n=0}^{M-1}E\{\hat{x}(n)-x(n)\}^2$ under the following assumptions.

1) The input $x(n)$ is assumed to be a zero mean real CWSS process.
2) The input $x(n)$ and the quantization noise $q(n)$ are uncorrelated processes, i.e., $E\{x(n)q(m)\} = 0 \quad \forall n, m$.
3) The quantization noise $q(n)$ is white with variance σ_q^2 as in (3).
4) The filters $P(e^{j\omega})$ and $V(e^{j\omega})$ are not constrained to be rational functions and can be non causal.
5) The average power spectral density $\hat{S}_{xx}(e^{j\omega})$ is positive for all ω. Furthermore, for the derivation of the optimum prefilter, we will also require $S_{xx}(e^{j\omega})$ and its first derivative to be continuous functions of frequency.

To solve the above problem, our approach will be the following. First, consider the single-band case of Fig. 8. Unlike previous quantization schemes, we observe that in the absence of the quantizer, the scheme of Fig. 8 is *not* a PR system. The error sequence $e(n) = \hat{x}(n) - x(n)$ has, in fact, two components: one due to the mismatch between the pre- and post-filters and the other due to the filtered quantization noise. We cannot therefore simply minimize the mean square reconstruction error as in the previous sections. Using the mse definition given above, we derive an expression for the average mean square reconstruction error $1/M\,\Sigma_{n=0}^{M-1}E\{e^2(n)\}$ in terms of the filters and the average power spectrum of the signal $x(n)$ and noise $q(n)$. The use of the average power

spectral density of the $(CWSS)_M$ input $x(n)$ in this case is not theoretically correct, even under the same quantizer assumptions as before. Nevertheless, it is necessary to work with this quantity to obtain any meaningful comparison between this more general setup and the one of the previous subsection. The calculus of variation is used as a tool to derive closed-form expressions for both the optimum prefilters and postfilters, which are then used to obtain the coding gain expression of Fig. 8. Finally, we will show how to the generalize the results for the multiband case of Fig. 20.

Theorem 7.2.1: For a fixed prefilter $P(e^{j\omega})$ and a given filter $F(e^{j\omega})$, the optimum postfilter $V_{\rm opt}(e^{j\omega})$ is as in (26), shown at the bottom of the page.

Proof: The average mean square reconstruction error can be expressed as follows:

$$\begin{aligned}\mathcal{E} &= \frac{1}{M}\sum_{n=0}^{M-1} E\{e^2(n)\} \\ &= \sigma_x^2 + \frac{1}{M}\sum_{n=0}^{M-1} E\{\hat{x}^2(n)\} - \frac{1}{M}\sum_{n=0}^{M-1} E\{\hat{x}(n)x(n)\} \\ &\quad - \frac{1}{M}\sum_{n=0}^{M-1} E\{x(n)\hat{x}(n)\} \\ &= \int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega})\frac{d\omega}{2\pi} + \frac{1}{M}\int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega})|P(e^{j\omega})|^2 \\ &\quad \cdot |V(e^{j\omega})|^2|F(e^{j\omega})|^2\frac{d\omega}{2\pi} + \frac{1}{M}\int_{-\pi}^{\pi} \sigma_q^2|V(e^{j\omega})|^2 \\ &\quad \cdot |F(e^{j\omega})|^2\frac{d\omega}{2\pi} - \frac{2}{M}\int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega})|F(e^{j\omega})|^2 \\ &\quad \cdot \Re\{P(e^{j\omega})V(e^{j\omega})\}\frac{d\omega}{2\pi} \end{aligned} \tag{27}$$

where $\Re$ stands for the real part. First, observe that the average mse dependency on the phase of the filters appears only in the last term. To minimize (27) with respect to the phase of the filters, the product $P(e^{j\omega})V(e^{j\omega})$ must be zero phase. To see this, simply set $P(e^{j\omega}) = |P(e^{j\omega})|e^{j\phi(\omega)}$ and $V(e^{j\omega}) = |V(e^{j\omega})|e^{j\Phi(\omega)}$. The real part of $P(e^{j\omega})V(e^{j\omega})$ is equal to $|V(e^{j\omega})|\,|P(e^{j\omega})|\cos(\phi(\omega)+\Phi(\omega))$. To minimize (27), $\cos(\phi(\omega)+\Phi(\omega))$ must be equal to one. Dropping the real notation $\Re$ in (27), we now turn to the magnitude squared response of the filters. We first fix the prefilter $P(e^{j\omega})$ and optimize $|V(e^{j\omega})|$. This can be done by applying the Euler–Lagrange equation from the calculus of variation theory [19] to (27). The resulting expression is (26). ■

It is interesting to note that the postfilter is independent of $F(e^{j\omega})$. Substituting (26) into (27), we obtain the following average mse expression in (28), shown at the bottom of the page.

Equation (28) is only a function of the magnitude squared response of the prefilter. From this point on, the problem under study is very similar to the one analyzed recently in [20], and, in fact becomes exactly the same by setting M and $F(e^{j\omega})$ to unity in (28). We will therefore omit the proofs of the upcoming theorems and refer to [20].

Theorem 7.2.2: The squared magnitude response $|P_{\rm opt}(e^{jw})|^2$ that minimizes $\mathcal{E}(|P|^2, b)$, which is given in (28), is also the solution of the following constrained optimization problem:

$$\begin{aligned}&\min_{|P(e^{jw})|^2}\int_{-\pi}^{\pi} \\ &\cdot \frac{\hat{S}_{xx}(e^{j\omega})(\hat{S}_{xx}(e^{j\omega})|P(e^{j\omega})|^2(1-|F(e^{jw})|^2)+c2^{-2b})}{\hat{S}_{xx}(e^{j\omega})|P(e^{jw})|^2+c2^{-2b}} \\ &\cdot \frac{d\omega}{2\pi}\end{aligned} \tag{29}$$

subject to

$$\int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega})|P(e^{j\omega})|^2\frac{d\omega}{2\pi} = 1. \tag{30}$$

Theorem 7.2.3: The prefilter $|P_{\rm opt}(e^{j\omega})|^2$ that minimizes (29) under the constraint (30) must have a magnitude response $|P_{\rm opt}(e^{j\omega})|^2$ in the form

$$\begin{aligned}&|P_{\rm opt}(e^{j\omega})|^2 \\ &\quad = \max\left(0, \frac{|F(e^{j\omega})|}{\sqrt{\hat{S}_{xx}(e^{j\omega})}}\left(\frac{1+c2^{-2b}}{\int_{-\pi}^{\pi}\sqrt{\hat{S}_{xx}(e^{j\omega})}|F(e^{j\omega})|\frac{d\omega}{2\pi}}\right.\right. \\ &\qquad \left.\left. - \frac{c2^{-2b}}{\sqrt{\hat{S}_{xx}(e^{j\omega})}}\right)\right) \quad \forall\omega\in[-\pi,\pi].\end{aligned} \tag{31}$$

$$V_{\rm opt}(e^{j\omega}) = \frac{1}{P(e^{j\omega})}\frac{\hat{S}_{xx}(e^{j\omega})}{\hat{S}_{xx}(e^{j\omega}) + \dfrac{c2^{-2b}}{|P(e^{j\omega})|^2}\displaystyle\int_{-\pi}^{\pi}\hat{S}_{xx}(e^{j\omega})|P(e^{j\omega})|^2\frac{d\omega}{2\pi}} \tag{26}$$

$$\mathcal{E}(|P|^2,b) = \int_{-\pi}^{\pi}\frac{\hat{S}_{xx}(e^{j\omega})\left(\hat{S}_{xx}(e^{j\omega})|P(e^{j\omega})|^2(M-|F(e^{jw})|^2)+c2^{-2b}M\displaystyle\int_{-\pi}^{\pi}\hat{S}_{xx}(e^{ju})|P(e^{ju})|^2\frac{du}{2\pi}\right)}{\hat{S}_{xx}(e^{j\omega})|P(e^{jw})|^2+c2^{-2b}\displaystyle\int_{-\pi}^{\pi}\hat{S}_{xx}(e^{ju})|P(e^{ju})|^2\frac{du}{2\pi}}\frac{d\omega}{2\pi}. \tag{28}$$

Theorem 7.2.4: With the optimal choice of prefilters and postfilters, the coding gain expression for the scheme of Fig. 8 is

$$\mathcal{G}_{\text{opt}} = (1 + c2^{-2b})M\mathcal{G}_{hw} \tag{32}$$

as long as $|P_{\text{opt}}(e^{j\omega})|^2$ in (31) is never set to zero $\forall\omega$. Here, $\mathcal{G}_{hw}$ is, again, the half whitening coding gain of the WSS process $y(n)$.

Note that in this case, the coding gain of the more general setup is a concatenation of three factors.

1) $\mathcal{G}_{hw}$ due to the noise shaping;
2) the oversampling factor M due to the signal model;
3) $1 + c2^{-2b}$ due to using a more general form of pre- and post filters.

To conclude this section, we would like to repeat the same procedure for the more general scheme of Fig. 20. We claim that for this case, the optimum postfilter is still given by (26), and the optimum prefilter magnitude squared response expression is obtained from (31) by simply replacing $|F(e^{j\omega})|$ with $\sqrt{\Sigma_{k=0}^{L-1}|F_k(e^{j\omega})|^2}$. To prove this, the key is to derive an expression for the average mean square reconstruction error of Fig. 20. Clearly, if we can show that $\mathcal{E}$ for the multiband case can be expressed as

$$\begin{aligned}\mathcal{E} &= E\{x^2(n)\} + \frac{1}{M}\sum_{n=0}^{M-1} E\{\hat{x}^2(n)\} \\ &\quad - \frac{1}{M}\sum_{n=0}^{M-1} E\{\hat{x}(n)x(n)\} - \frac{1}{M}\sum_{n=0}^{M-1} E\{x(n)\hat{x}(n)\} \\ &= \int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega})\frac{d\omega}{2\pi} + \frac{1}{M}\int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega})|P(e^{j\omega})|^2 \\ &\quad \cdot |V(e^{j\omega})|^2 \sum_{k=0}^{L-1} |F_k(e^{j\omega})|^2 \frac{d\omega}{2\pi} + \frac{1}{M}\int_{-\pi}^{\pi} \sigma_q^2 |V(e^{j\omega})|^2 \\ &\quad \cdot \sum_{k=0}^{L-1} |F(e^{j\omega})|^2 \frac{d\omega}{2\pi} - \frac{2}{M}\Re\int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega})P(e^{j\omega}) \\ &\quad \cdot V(e^{j\omega}) \sum_{k=0}^{L-1} |F(e^{j\omega})|^2 \frac{d\omega}{2\pi}\end{aligned} \tag{33}$$

then from the previous analysis, the above claim follows immediately. To derive (33), we need to only consider the second term and one of the cross terms. The second term $1/M\,\Sigma_{n=0}^{M-1} E\{\hat{x}^2(n)\}$ is the variance of the signal estimate at the output of Fig. 20, but from Result 2 of Section III, we know that it is equal to $1/M\,\Sigma_{k=0}^{L-1}\sigma_{y_k}^2$, where $\sigma_{y_k}^2$ is the variance of the signal estimate before the kth channel downsampler $k = 0, \cdots, L-1$. Substituting with $\sigma_{y_k}^2$ in this last relation, we obtain the second and third integral in (33). Consider now one of the cross terms, say, $1/M\,\Sigma_{n=0}^{M-1} E\{\hat{x}(n)x(n)\}$. We can rewrite $\hat{x}(n)$ as $\Sigma_{k=0}^{L-1}\hat{x_k}(n)$, where $\hat{x_k}(n)$ is the signal estimate at the output of the kth channel. By the linearity of the expectation, this gives $1/M\,\Sigma_{k=0}^{L-1}\Sigma_{n=0}^{M-1} E\{\hat{x_k}(n)x(n)\}$. By interpreting the single-band case as the kth channel, the last integral follows easily. Equation (33) is therefore established and the claim is proved.

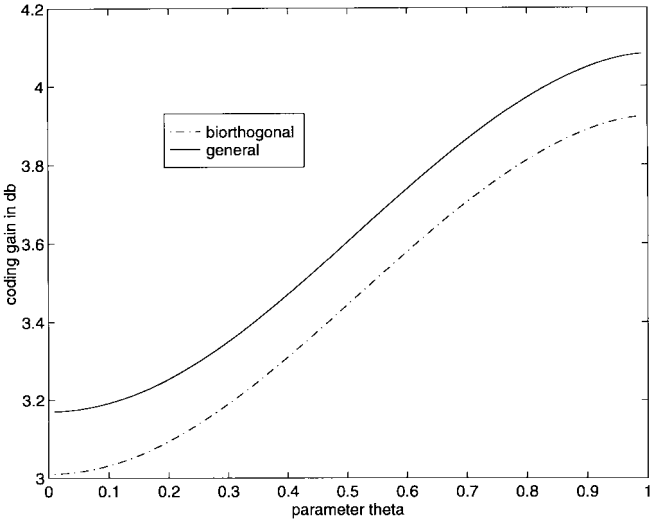


Fig. 21. Coding gain curves for the MA(1) case with $b = 3$ and $c = 2.4$.

Example 7.1—Case of a MA(1) Process $y(n)$: Assume that the input $x(n)$ is modeled as in Fig. 1 with $M = 2$ and $F(z) = (1/\sqrt{2})(1+z^{-1})$. Let the driving WSS signal $y(n)$ be a zero mean Gaussian MA(1) process with an autocorrelation sequence in the form

$$R_{yy}(k) = \begin{cases} 1, & k = 0. \\ \dfrac{\theta}{1+\theta^2}, & k = 1, -1 \\ 0, & \text{otherwise.} \end{cases}$$

The MA(1) process has to have $(|R_{yy}(1)|/R_{yy}(0)) \le 1/2$ to ensure that the power spectral density is indeed non-negative. We therefore restrict θ to be between -1 and 1. The power spectrum of the MA(1) process is given by

$$S_{yy}(e^{j\omega}) = 1 - 2\frac{\theta}{(1+\theta^2)}\cos(\omega). \tag{34}$$

Substituting (34) in (21), the coding gain expression of the scheme Fig. 6 becomes

$$\mathcal{G}_{\text{opt}} = \frac{2(1+\theta^2)}{\left(\displaystyle\int_{-\pi}^{\pi}\sqrt{(1+\theta^2-2\theta\cos(\omega))}\frac{d\omega}{2\pi}\right)^2}. \tag{35}$$

The integral in (35) is equal to $F(-0.5, -0.5; 1; \theta^2)$, where $F(a, b; c; d)$ is Gauss's hypergeometric function. From [21], $F(-0.5, -0.5; 1; \theta^2)$ can be rewritten as $(1 + \theta)F(-0.5, 0.5; 1; 4\theta/(1+\theta)^2)$. This, in turn, can be simplified to $(1+\theta)(2/\pi)E(2\sqrt{(|\theta|)}/(1+\theta))$, where $E(\cdot)$ is the complete elliptic integral of the second kind. The coding gain of the more general system can be obtained by multiplying (35) by $(1 + c2^{-2b})$ and obviously depends on the number of bits b. The plots of the coding gain are illustrated in Fig. 21 for $b = 3$ and $c = 2.4$.

Example 7.2—Case of an AR(1) Process $y(n)$: With the same assumptions as in Example 7.1, let the driving signal $y(n)$ be a zero mean Gaussian AR(1) process with an autocorrelation sequence in the form $R_{yy}(k) = \rho^{|k|}$, where ρ is

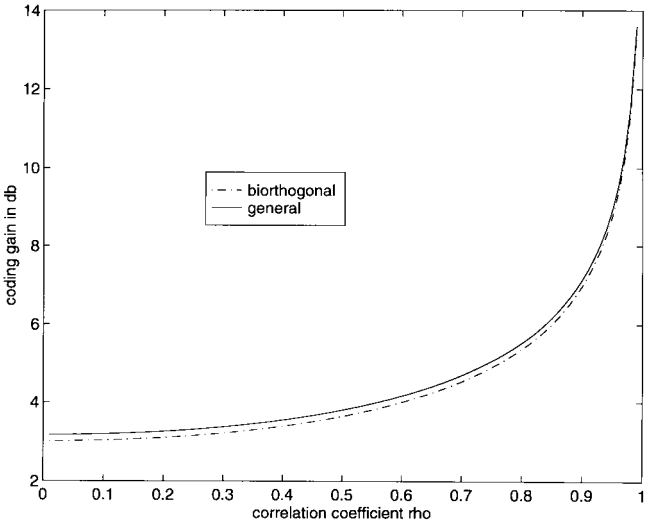


Fig. 22. Coding gain curves for the AR(1) case with $b = 3$ and $c = 2.4$.

between 0 and 1. The power spectrum of the AR(1) process is

$$S_{yy}(e^{j\omega}) = \frac{1-\rho^2}{1+\rho^2-2\rho\cos(\omega)}. \tag{36}$$

Substituting (36) in (21), the coding gain expression for the scheme of Fig. 6 is

$$\mathcal{G}_{\text{opt}} = \frac{2}{(1-\rho^2)\left(\displaystyle\int_{-\pi}^{\pi} \frac{1}{\sqrt{(1+\rho^2-2\rho\cos(\omega))}}\frac{d\omega}{2\pi}\right)^2}. \tag{37}$$

The integral in (37) is equal to $(2/\pi)K(\rho)$, where $K(\rho)$ is the complete elliptic integral of the first kind [21]. Again, the coding gain of the more general system is obtained by multiplying (37) by $(1+c2^{-2b})$. The plots of the coding gain are shown in Fig. 22 for $b = 3$ and $c = 2.4$.

VIII. Noise Shaping by $(LPTV)_M$ Pre- and Post-filters

In this section, we consider using $(LPTV)_M$ pre- and post-filters instead of LTI ones surrounding a periodically time-varying $((PTV)_M)$ quantizer. Since the signal model $x(n)$ is $(CWSS)_M$, restricting ourselves to linear time-invariant noise shaping filters and quantizers is a loss of generality. Any optimum configuration for such processes should consist of $(LPTV)_M$ filters surrounding a $((PTV)_M)$ quantizer. Using some well-known multirate results, it can be shown that this new quantization configuration is equivalent to an M-channel maximally decimated filter bank with M subband quantizers [5]. We will further impose the PR condition in the absence of quantization by confining ourselves to the class of PR filter banks. It follows that $\boldsymbol{R}(e^{j\omega}) = \boldsymbol{E}^{-1}(e^{j\omega})$, where $\boldsymbol{E}(e^{j\omega})$ and $\boldsymbol{R}(e^{j\omega})$ denote, respectively, the analysis and synthesis polyphase matrices [5]. Equivalently, the analysis and synthesis filters satisfy the biorthogonality condition $(P_k(e^{j\omega})Q_m(e^{j\omega}))|\downarrow_M = \delta(m-k)$ for all k, m. The goal is then to find the set of M analysis and synthesis filters $P_k(e^{j\omega})$ and $Q_k(e^{j\omega})$ (equivalently the analysis and synthesis polyphase matrices) that minimize the average mean square error at the output due to the quantization noise. Because the general $(LPTV)_M$ problem is difficult to track analytically, we will only study two special forms of the above setup. The first case assumes that $\boldsymbol{E}(e^{j\omega})$ is diagonal with diagonal elements equal to $V_k(e^{j\omega})$. It follows that $\boldsymbol{R}(e^{j\omega})$ is also diagonal with diagonal elements equal to $(1/V_k(e^{j\omega}))$ for each k. The second case assumes that $\boldsymbol{E}(e^{j\omega})$ is paraunitary, and we choose $\boldsymbol{R}(e^{j\omega}) = \boldsymbol{E}^{\dagger}(e^{j\omega})$. Alternatively, the synthesis filters $Q_k(e^{j\omega})$ are equal to $\tilde{P}_k(e^{j\omega})$ for each k and $(P_k(e^{j\omega})\tilde{P}_m(e^{j\omega}))|\downarrow_M = \delta(m-k)$ for all k, m. These two special forms are intermediate between one extreme (the LTI case) and the other (the general $(LPTV)_M$ case).

A. Letting the Synthesis Filter Be the Inverse of the Analysis Filter

Let $\boldsymbol{E}(e^{j\omega})$ be a diagonal matrix with diagonal elements equal to $V_k(e^{j\omega})$ and $\boldsymbol{R}(e^{j\omega})$ be also diagonal with diagonal elements equal to $(1/V_k(e^{j\omega}))$ for each k. The quantization configuration is shown in Fig. 9 for the single-band case and Fig. 23 for the multiband case. The scalar quantizers labeled $\mathcal{Q}$ are modeled as additive noise sources $q_k(n)$ and individually satisfy relation (3). Throughout this section, we will assume that the subband quantization noise sources $q_k(n)$ are white and pairwise uncorrelated, i.e., the noise power spectral density matrix is given by

$$\boldsymbol{S}_{qq}(e^{j\omega}) = \begin{pmatrix} \sigma_{q_0}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{q_1}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma_{q_{M-1}}^2 \end{pmatrix}. \tag{38}$$

The goal is then to jointly allocate the subband bits b_k under a fixed bit rate

$$b = \frac{1}{M}\sum_{k=0}^{M-1} b_k \tag{39}$$

and optimize $V_k(e^{j\omega})$ in order to minimize the average mse at the output of Figs. 9 and 23. Our strategy is as follows: We first find the optimum solution for the single-band case of Fig. 9. Then, by interpreting the single-band model as one of the L channels of the more general multiband case, the optimum solution for Fig. 23 follows.

Theorem 8.1.1: Consider the scheme of Fig. 9 under the above assumptions. The optimum filter $V_{\text{opt}}(e^{j\omega})$ that minimizes the average mean square reconstruction error at the output is *independent of* k and has the magnitude squared response

$$|V_{\text{opt}}(e^{j\omega})|^2 = \frac{1}{\sqrt{S_{yy}(e^{j\omega})}} \tag{40}$$

where $S_{yy}(e^{j\omega})$ is the power spectrum of the WSS process $y(n)$ in Fig. 1. With the above optimum filter expression, the coding gain of Fig. 9 is then given by

$$\mathcal{G}_{\text{opt}} = \frac{\sigma_y^2}{M\left(\displaystyle\prod_{k=0}^{M-1}\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}|\tilde{R}_k(e^{j\omega})|^2\frac{d\omega}{2\pi}\right)^{2/M}} \tag{41}$$

where $\tilde{R}_k(e^{j\omega})$ is the kth polyphase component of $\tilde{F}(e^{j\omega})$.

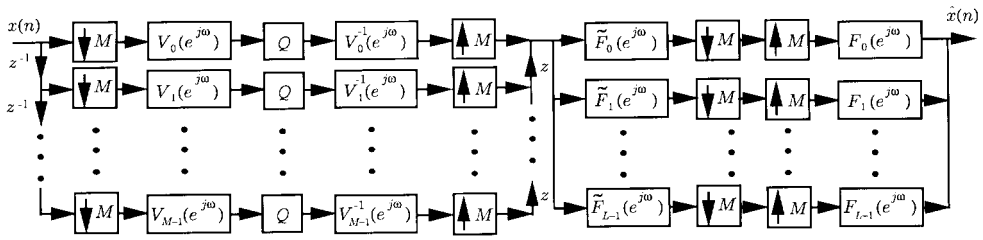


Fig. 23. Scheme 1 for noise shaping using $(LPTV)_M$ pre- and post-filters (the multiband case).

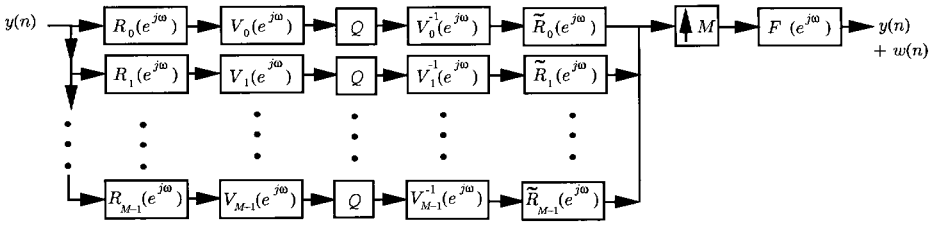


Fig. 24. Equivalent representation of Fig. 9.

Proof: Since the system has the PR property in the absence of quantization, the error $e(n)$ at the output is simply the filtered quantization noise signal. After the downsampler, the filtered noise component $w(n)$ is WSS. By Result 2 of Section III, $\mathcal{E} = (1/M)\sigma_w^2$. To compute σ_w^2, we express the filter $\tilde{F}(e^{j\omega})$ in terms of its M polyphase components $\tilde{R}_k(e^{j\omega})$. Because the input signal $x(n)$ is modeled as in Fig. 1, we can also invoke the polyphase identity (see [5, p. 133]) at the input to simplify Fig. 9 (see Fig. 24). The interpolation filter was not drawn because we are really interested in evaluating σ_w^2 rather than σ_e^2. Since the quantization noise sources are assumed to be white and uncorrelated, the average mean squared error is therefore given by

$$\begin{aligned}\mathcal{E} &= \frac{c}{M}\sum_{k=0}^{M-1} 2^{-2b_k}\sigma_{y_k}^2 \int_{-\pi}^{\pi} \frac{|\tilde{R}_k(e^{j\omega})|^2}{|V_k(e^{j\omega})|^2}\frac{d\omega}{2\pi} \\ &= \frac{c}{M}\sum_{k=0}^{M-1} 2^{-2b_k}\int_{-\pi}^{\pi} S_{yy}(e^{j\omega})|R_k(e^{j\omega})|^2|V_k(e^{j\omega})|^2 \\ &\quad \cdot \frac{d\omega}{2\pi}\int_{-\pi}^{\pi}\frac{|\tilde{R}_k(e^{j\omega})|^2}{|V_k(e^{j\omega})|^2}\frac{d\omega}{2\pi}. \end{aligned} \tag{42}$$

Using the AM–GM inequality (39) and the fact that $|R_k(e^{j\omega})|^2 = |\tilde{R}_k(e^{j\omega})|^2$, (42) reduces to

$$\begin{aligned}\mathcal{E} \geq c2^{-2b}\Bigg(\prod_{i=0}^{M-1}\int_{-\pi}^{\pi} S_{yy}(e^{j\omega})|V_k(e^{j\omega})|^2|\tilde{R}_k(e^{j\omega})|^2 \\ \cdot \frac{d\omega}{2\pi}\int_{-\pi}^{\pi}\frac{|\tilde{R}_k(e^{j\omega})|^2}{|V_k(e^{j\omega})|^2}\frac{d\omega}{2\pi}\Bigg)^{1/M}. \end{aligned} \tag{43}$$

Applying the Cauchy–Schwartz inequality to each term in (43), we get

$$\mathcal{E}_{\min} = c2^{-2b}\left(\prod_{k=0}^{M-1}\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}|\tilde{R}_k(e^{j\omega})|^2\frac{d\omega}{2\pi}\right)^{2/M}. \tag{44}$$

This minimum bound is achieved by choosing $|V_{\rm opt}(e^{j\omega})|^2$ as in (40). Finally, (41) follows immediately from the definition of the coding gain (3) and the fact that $\sigma_x^2 = \sigma_y^2/M$. ■

The LTI Case is Indeed a Loss of Generality: Since the class of $(LPTV)_M$ filters and $(PTV)_M$ quantizers include the LTI case, it is clear that the performance of this more general class of filters and quantizers is at least as good as the LTI one. We have already shown that the optimum $(LPTV)_M$ filter for Fig. 9 reduces to an LTI one. The question then becomes the following: Is the $(PTV)_M$ quantizer providing any excess gain over the LTI case, and if so, by how much? We show next that even in this restricted form of $(LPTV)_M$ filters, the coding gain of the above scheme is always greater than the LTI one, except when the magnitude squared response of the polyphase components $R_k(e^{j\omega})$ of $F(e^{j\omega})$ are equal for all k. Starting from the denominator of (22) (the coding gain expression of Fig. 6), we can write the following series of steps:

$$\begin{aligned}&\frac{1}{M}\left(\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}\frac{d\omega}{2\pi}\right)^2 \\ &\quad = \frac{1}{M}\left(\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}\sum_{k=0}^{M-1}|\tilde{R}_k(e^{j\omega})|^2\frac{d\omega}{2\pi}\right)^2 \\ &\quad = \frac{1}{M}\left(\sum_{k=0}^{M-1}\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}|\tilde{R}_k(e^{j\omega})|^2\frac{d\omega}{2\pi}\right)^2 \\ &\quad \geq \frac{1}{M}\times M^2\left(\left(\prod_{i=0}^{M-1}\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}|\tilde{R}_k(e^{j\omega})|^2\frac{d\omega}{2\pi}\right)^{1/M}\right)^2 \\ &\quad = M\left(\prod_{i=0}^{M-1}\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}|\tilde{R}_k(e^{j\omega})|^2\frac{d\omega}{2\pi}\right)^{2/M} \end{aligned} \tag{45}$$

where the last line in (45) is the denominator of (41). Since the numerator is the same in both cases, the claim is proved. The first equality in (45) is obtained by using the power complementary property of the polyphase components of

$F(e^{j\omega})$. The second line is a consequence of the linearity of the integral. The third line results from applying the AM–GM inequality. From the AM–GM formula, we know that equality is achieved if and only if all $|\tilde{R}_k(e^{j\omega})|^2$ are equal. From Fig. 24, we can see that this makes perfect sense. If all $|\tilde{R}_k(e^{j\omega})|^2$ are equal and since the optimum filters $V_k(e^{j\omega})$ are independent of k, the variance of the subband quantizer inputs will be all equal. There is therefore no variance disparity in the subbands, and optimum bit allocation of the subband quantizers (which depends on the AM–GM inequality) cannot produce any gain. Using the single band result, we can now derive closed-form expressions for the optimum $V_{\mathrm{opt}_k}(e^{j\omega})$ and the average minimum mean squared error for the multiband case.

Theorem 8.1.2: Consider the scheme of Fig. 23 under the above assumptions. The optimum filter $V_{\mathrm{opt}_k}(e^{j\omega})$ (for each k) that minimizes the average mean square reconstruction error at the output has the magnitude squared response

$$|V_{\mathrm{opt}_k}(e^{j\omega})|^2 = \frac{\sqrt{\sum_{i=0}^{L-1}|\tilde{R}_{ik}(e^{j\omega})|^2}}{\sqrt{S_k(e^{j\omega})}} \tag{46}$$

where $\tilde{R}_{ik}(e^{j\omega})$ is the kth polyphase component of the ith filter $\tilde{F}_i(e^{j\omega})$, and $S_k(e^{j\omega}) = \Sigma_{i=0}^{L-1} S_{y_i}(e^{j\omega})|\tilde{R}_{ik}(e^{j\omega})|^2$ is the power spectrum of kth channel. Using the above optimum filters, the coding gain of Fig. 23 is then given by

$$\mathcal{G}_{\mathrm{opt}} = \frac{\sigma_y^2}{M\left(\prod_{k=0}^{M-1}\int_{-\pi}^{\pi}\sqrt{S_k(e^{j\omega})}\sqrt{\sum_{i=0}^{L-1}|\tilde{R}_{ik}(e^{j\omega})|^2}\frac{d\omega}{2\pi}\right)^{2/M}}. \tag{47}$$

Proof: By interpreting the single-band result as one of the L channels of the multiband model and by using Result 2 of Section III, the average mse can be expressed as

$$\begin{aligned}\mathcal{E} &= \frac{c}{M}\sum_{k=0}^{M-1} 2^{-2b_k}\sigma_{y_k}^2\int_{-\pi}^{\pi}\frac{\sum_{i=0}^{L-1}|\tilde{R}_{ik}(e^{j\omega})|^2}{|V_k(e^{j\omega})|^2}\frac{d\omega}{2\pi}\\ &= \frac{c}{M}\sum_{k=0}^{M-1} 2^{-2b_k}\int_{-\pi}^{\pi} S_k(e^{j\omega})|V_k(e^{j\omega})|^2\frac{d\omega}{2\pi}\\ &\quad\cdot\int_{-\pi}^{\pi}\frac{\sum_{i=0}^{L-1}|\tilde{R}_{ik}(e^{j\omega})|^2}{|V_k(e^{j\omega})|^2}\frac{d\omega}{2\pi}.\end{aligned} \tag{48}$$

Using the same inequalities as in the proof of Theorem 8.1.1, we can immediately derive (46) and (47). ∎

Following the same type of reasoning as before, we again expect the coding gain of the more general $(LPTV)_M$ case of Fig. 23 to be higher than the analogous LTI one of Fig. 19. However, the complexity of the expressions (25) and (47) in this case prevents a formal mathematical proof.

Example 8.1—Equal Polyphase Components: Assume that the input $x(n)$ is modeled as in Fig. 1, where the upsampler $M = 2$ and the driving input $y(n)$ is a zero mean Gaussian AR(1) process with correlation coefficient $0<\rho<1$. Furthermore, let $F(z)$ be the optimum FIR compaction filter of length two given by $(1/\sqrt{2})(1+z^{-1})$. The filter actually corresponds to one of the channels of a 2×2 KLT that is independent of the input statistics. In this case, the polyphase components of $F(e^{j\omega})$ are $R_0(e^{j\omega}) = R_1(e^{j\omega}) = (1/\sqrt{2})$. Substituting in (41) and simplifying, we get (21), which is the coding gain expression of Fig. 6. In Example 7.2., a closed-form expression was derived for the AR(1) case, and a plot of the coding gain is shown in Fig. 22.

Example 8.2—Unequal Polyphase Components: With the same set of assumptions of Example 8.1, let the filter $F(z)$ be the optimum FIR compaction filter of length four. With $M = 2$ and assuming an AR(1) process, the following closed-form expression was derived in [22] for the optimum compaction filter:

$$F(z) = a + cz^{-1} + bz^{-2} + dz^{-3} \tag{49}$$

where

$$\begin{aligned}a &= \frac{1}{p\sqrt{2}}\sqrt{\sqrt{p}+\sqrt{q}},\\ b &= \frac{1}{p\sqrt{2}}\left(\sqrt{\sqrt{p}+\sqrt{q}} - \sqrt{p}\sqrt{\sqrt{p}-\sqrt{q}}\right)\\ c &= \frac{1}{p\sqrt{2}}\left(\sqrt{p}\sqrt{\sqrt{p}+\sqrt{q}} - \sqrt{\sqrt{p}-\sqrt{q}}\right)\\ d &= -\frac{1}{p\sqrt{2}}\sqrt{\sqrt{p}-\sqrt{q}}\end{aligned}$$

and $p = 3+\rho^2, q = 2+\rho^2$. The polyphase components of $F(e^{j\omega})$ are $R_0(e^{j\omega}) = a + be^{-j\omega}$ and $R_1(e^{j\omega}) = c + de^{-j\omega}$. Substituting the power spectrum expression of an AR(1) process given by (36) into (41) and using some useful integral formulas (see [21, p. 429]), we can derive the coding gain expression in (50), shown at the bottom of the next page, for the scheme of Fig. 9, where $K(\cdot)$ is the complete elliptic integral of the first kind, and $E(\cdot)$ is the complete elliptic integral of the second kind. There is a reason for writing the denominator of (50) in this form. It can be shown that the factors $(2/\pi)((a^2+b^2+(2ab/\rho))K(\rho) - (2ab/\rho)E(\rho))$ and $(2/\pi)((c^2+d^2+(2cd/\rho))K(\rho) - (2cd/\rho)E(\rho))$ represent the variance of the outputs $R_0(e^{j\omega})$ and $R_1(e^{j\omega})$, respectively [with an input with power spectrum $\sqrt{S_{yy}}(e^{j\omega})$]. Their product is the geometric mean that produces the extra gain over the LTI case. The further away they are in magnitude, the more gain we will obtain. The plots of the coding gain formulas (37) and (50) are shown in Fig. 25. We notice that the coding gain of the $(LPTV)_M$ case is indeed greater than the LTI one for all values of ρ, although not by a substantial amount for the AR(1) process $y(n)$.

B. Using an Orthonormal FB

Consider now the M-channel orthonormal FB shown in Fig. 10 for the single-band model and in Fig. 26 for the multiband model. As in the previous subsection, we first analyze

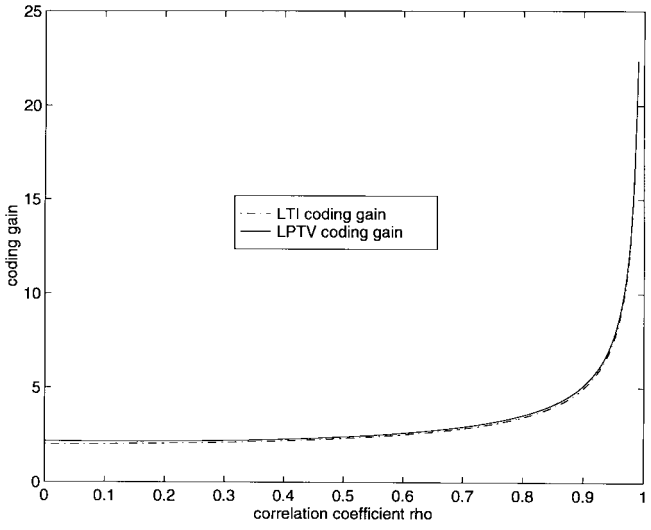


Fig. 25. Coding gain curves of the LTI and $(LPTV)_M$ cases under the assumption of a single-band model with $M = 2$ and $y(n)$ is an AR(1) process.

the single-band case in detail and then use the corresponding results to derive analogous expressions for the multiband case. The quantization noise assumptions of the previous subsection are still true here. The goal is again to jointly allocate the subband bits b_k under the constraint (39) and optimize the orthonormal filter bank in order to minimize the average mse.

Theorem 8.2.1: Consider the scheme of Fig. 10 under the above assumptions. The synthesis section of the optimum orthonormal FB $\{P_k(e^{j\omega})\}$ corresponds to choosing one of the filters, say, $\tilde{P}_0(e^{j\omega})$, to be equal to $\tilde{F(e^{j\omega})}$ and the remaining filters $\tilde{P}_k(e^{j\omega}), k = 1, \cdots, M-1$, to be orthogonal to $\tilde{P}_0(e^{j\omega})$. In this case, the optimum orthonormal FB reduces to Fig. 5, where the quantizer $\mathcal{Q}$ is allocated Mb bits according to (37).

Proof: By applying the blocking operation and using the polyphase representation [5], the scheme of Fig. 10 can be redrawn as in Fig. 27, where $\boldsymbol{E}(e^{j\omega})$ is the polyphase matrix of the analysis bank, $\boldsymbol{E}^{\dagger}(e^{j\omega})$ is the polyphase matrix of the synthesis bank, and $\tilde{R}_k(e^{j\omega}), k = 0, \cdots, M-1$ are the M polyphase components of the filter $\tilde{F}(e^{j\omega})$. Let $\boldsymbol{U}(e^{j\omega})$ be the $1 \times M$ vector whose kth element is $\tilde{R}_k(e^{j\omega})$. Then, the average mse can be expressed as

$$\mathcal{E} = \frac{1}{M}\int_{-\pi}^{\pi} \text{trace}\,(\boldsymbol{U}(e^{j\omega})\boldsymbol{E}(e^{j\omega})\boldsymbol{S_{qq}}\boldsymbol{E}^{\dagger}(e^{j\omega})\boldsymbol{U}^{\dagger}(e^{j\omega}))\frac{d\omega}{2\pi}. \tag{51}$$

Since the integrand is in a quadratic form, the trace operator can be removed. Furthermore, since $\boldsymbol{E}(e^{j\omega})\boldsymbol{E}^{\dagger}(e^{j\omega}) = I$ by orthonormality and $\boldsymbol{U}(e^{j\omega})\boldsymbol{U}^{\dagger}(e^{j\omega}) = 1$ by the Nyquist property of the $F(e^{j\omega})$, we can rewrite (51) as

$$\mathcal{E} = \frac{1}{M}\int_{-\pi}^{\pi} \frac{\boldsymbol{P}(e^{j\omega})\boldsymbol{S_{qq}}\boldsymbol{P}^{\dagger}(e^{j\omega})}{\boldsymbol{P}(e^{j\omega})\boldsymbol{P}^{\dagger}(e^{j\omega})}\frac{d\omega}{2\pi} \tag{52}$$

where $\boldsymbol{P}(e^{j\omega}) = \boldsymbol{U}(e^{j\omega})\boldsymbol{E}(e^{j\omega})$. Since the integrand of (52) is positive for all ω, minimizing (52) is equivalent to minimizing the integrand at each frequency. However, for any fixed frequency ω_0, the ratio $\boldsymbol{P}(e^{j\omega_0})\boldsymbol{S_{qq}}\boldsymbol{P}^{\dagger}(e^{j\omega_0})/\boldsymbol{P}(e^{j\omega_0})\boldsymbol{P}^{\dagger}(e^{j\omega_0}))$ is a Rayleigh quotient. For each frequency ω, the minimizing vector $\boldsymbol{P}_{\text{opt}}(e^{j\omega})$ has the form $(0 \quad \cdots \quad 1 \quad \cdots \quad 0)$, where the 1 in the ith position corresponds to the minimum noise variance $\sigma_{q_i}^2$. Since $\boldsymbol{P}(e^{j\omega}) = \boldsymbol{U}(e^{j\omega})\boldsymbol{E}(e^{j\omega})$, the minimizing vector $\boldsymbol{P}_{\text{opt}}(e^{j\omega})$ can be obtained by setting the ith column in $\boldsymbol{E}(e^{j\omega})$ to be equal to $\boldsymbol{U}^{\dagger}(e^{j\omega})$ and all the remaining columns to be orthogonal to $\boldsymbol{U}(e^{j\omega})$. This is equivalent to the statement of the theorem. ■

The optimum orthonormal filter bank thus reduces to the scheme of Fig. 5 with Mb bits allocated to the quantizer. The result of Theorem 8.2.1 is very intuitive and somehow expected: Filter and decimate the oversampled signal $x(n)$ according to its model and then quantize $y(n)$ in Fig. 5 with $\hat{b} = Mb$ bits/sample. As we mentioned before, this amounts to fixing the bit rate (number of bits per second) in order to trade quantization resolution with sampling rate. It is interesting, though, to see that this very intuitive scheme is equivalent to using an optimum orthonormal FB as a sophisticated quantizer to the input $x(n)$. With (3) in mind, the coding gain expression can be derived following the lines of the proof of Theorem 5.1 and is equal to $2^{2b(M-1)}$. This is an exponential gain that can be quite large for moderate values of M but, unlike all previous schemes, depends on the bit rate b. Finally, to end this section, we would like to derive an analogous result (to Theorem 8.2.1) for the multiband case.

Theorem 8.2.2: Consider the scheme of Fig. 26 under the same assumptions. The synthesis section of the optimum orthonormal FB corresponds to choosing L of the filters to be equal to $\tilde{F}_k(e^{j\omega})$ and the remaining filters $Q_k(e^{j\omega}), k = L+1, \cdots, M-1$ to be the $M-L-1$ orthogonal filters to $F_i(e^{j\omega}), i = 0, \cdots, L-1$. In this case, the optimum orthonormal FB reduces to Fig. 18 with an equivalent average number of bits $\hat{b}$ equal to Mb/L bits.

Proof: By interpreting the single-band result as one of the L channels of the multiband model and by using Result 2 and (39), the result follows immediately. ■

With the above $\hat{b}$, we can now perform an optimum allocation of subband bits for the scheme of Fig. 18. This is a standard allocation problem that arises in subband coding application [15]. By applying the AM–GM inequality to the

$$\mathcal{G}_{\text{opt}} = \frac{1}{2(1-\rho^2)\frac{2}{\pi}\left(\left(a^2+b^2+\frac{2ab}{\rho}\right)K(\rho) - \frac{2ab}{\rho}E(\rho)\right)\frac{2}{\pi}\left(\left(c^2+d^2+\frac{2cd}{\rho}\right)K(\rho) - \frac{2cd}{\rho}E(\rho)\right)} \tag{50}$$

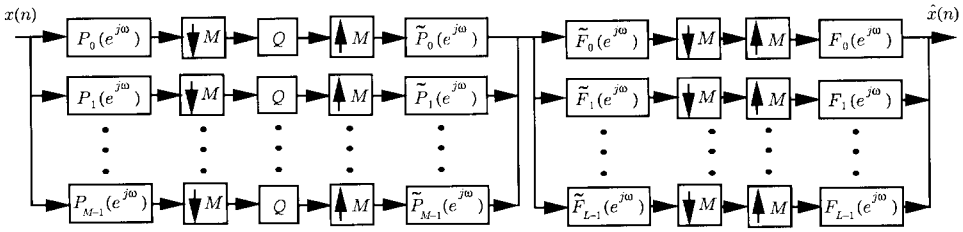


Fig. 26. Scheme 2 for noise shaping using $(LPTV)_M$ pre- and post-filters (the multiband case).

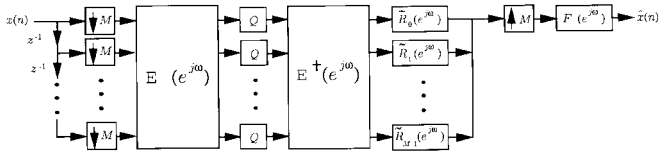


Fig. 27. Polyphase representation of Fig. 10.

output error expression $\mathcal{E} = c(1/M)\,\Sigma_{k=0}^{L-1}\,2^{-2b_k}\sigma_{y_k}^2$, we get

$$\mathcal{E}_{\min} = c2^{-2b(M/L)}\frac{L}{M}\left(\prod_{i=0}^{L-1}\sigma_{y_i}^2\right)^{1/L} \tag{53}$$

which can be achieved by setting $b_k = b + 0.5\log_2\sigma_{y_k}^2 - 0.5\log_2\,\Pi_{i=0}^{L-1}\,(\sigma_{y_i}^2)^{1/L}$. This optimum bit allocation formula will in almost all cases yield noninteger solution for the bits. A quick remedy might be to use a simple rounding procedure or a more sophisticated algorithm [23] to obtain integer solutions. A detailed discussion of the topic of allocating *integer* bits to the channel quantizers is, however, outside the scope of the paper. The noise variance in Fig. 12 simplifies to $c2^{-2b}(L/M)((1/L)\,\Sigma_{k=0}^{L-1}\,\sigma_{y_k}^2)$. The coding gain expression takes, therefore, the form

$$\mathcal{G}_{\text{opt}} = 2^{2b((M/L)-1)}\frac{AM(\sigma_{y_i}^2)}{GM(\sigma_{y_i}^2)} \tag{54}$$

where

AM arithmetic mean;
GM geometric mean;
$\sigma_{y_i}^2$ variance of the ith signal $y_i(n)$ in Fig. 2.

We observe that when $L = 1$, we get the coding gain of the single-band case, and when $L = M$, the scheme of Fig. 18 reduces to an orthonormal FB, the average number of bits is equal to b, and (54) reduces to the well-known expression of the coding gain of an orthonormal FB.

IX. Concluding Remarks

To complete the work in this paper, two important problems remain open and are potential ground for future research. First, the most general $(LPTV)_M$ scheme has not been studied in this paper. The optimization of such a scheme is closely related to another challenging problem, namely, the optimization of a biorthogonal FB according to the input signal statistics. Second, the theory and design of an optimal M-channel PCFB requires further investigation. Partial results have been obtained in [14], but the problem, in its full generality, remains open. For example, we can ask the following questions: What are the conditions that can guarantee the existence of a PCFB, and how can we achieve them? If these conditions are satisfied, how do we find the optimal PCFB? Answering these type of questions will help finding the multiband model of Fig. 2.

Appendix A

Proof of Result 1 in Section IV: The interpolated subband signals can be expressed as $x_i(n) = \Sigma_k\, y_i(k)f_i(n - Mk)$. Hence

$$\begin{aligned} &E[x_i(n)x_j^*(n - Mm)] \\ &\quad = E\left[\sum_{k'} y_i(k')f_i(n - Mk')\sum_{k} y_j^*(k)\right. \\ &\qquad \left.\cdot\, f_j^*(n - Mk - Mm)\right]. \end{aligned} \tag{55}$$

Let $r(u)$ be the cross correlation between the jointly WSS processes $y_i(n)$ and $y_j(n)$, that is, $r(u) = E[y_i(n)y_j^*(n - u)$. Using the change of variable $k'-k = l$, the preceding equation becomes

$$\begin{aligned} &E[x_i(n)x_j^*(n - Mm)] \\ &= \sum_{l} r(l)\sum_{k'} f_i(n - Mk')f_j^*(n + M(l - m) - Mk'). \end{aligned} \tag{56}$$

Substituting (56) in the left-hand side of (4), we get

$$\begin{aligned} &\frac{1}{M}\sum_{l} r(l)\sum_{k'}\sum_{k=0}^{M-1} f_i(n - (Mk' + k))f_j^*(n + M(l - m) \\ &\qquad - (Mk' + k)). \end{aligned} \tag{57}$$

Since M is positive, k' and k are integers, and $0 \le k < M$, we can always replace $Mk' + k$ by an integer u. That is, there always exist an integer u such that k' is the quotient, and k is the remainder obtained from dividing u by M. We can therefore rewrite (57) as

$$\begin{aligned} &\frac{1}{M}\sum_{l} r(l)\sum_{u} f_i(n - u)f_j^*(n + M(l - m) - u) \\ &\qquad = \frac{1}{M}\sum_{l} r(l)\sum_{k} f_i(k)f_j^*(k + M(l - m)). \end{aligned} \tag{58}$$

However, the orthonormality of the FB implies, in particular, that $\Sigma_k\, f_i(k)f_j^*k + M(l - m)) = 0 \quad \forall l, m$. Thus, the inner sum in (58) reduces to zero, and the result follows. ■

APPENDIX B

Phase Randomization of a $(CWSS)_M$ Process: A WSS process $\hat{x}(n)$ can be obtained from a $(CWSS)_M$ process $x(n)$ by introducing a random shift θ in the $(CWSS)_M$ signal $x(n)$ [16]–[18]. The parameter θ is a discrete random variable that can take any integer value from 0 to $M-1$ with equal probability $1/M$. Furthermore, the random variable θ is assumed to be independent of $x(n)$. The autocorrelation function of $\hat{x}(n)$ is given by

$$\begin{aligned} R_{\hat{x}\hat{x}}(n,k) &= E\{\hat{x}(n)\hat{x}(n-k)\} \\ &= E_\theta\{E\{x(n-\theta)x(n-k-\theta)|\theta\}\} \\ &= E_\theta\{R_{yy}(n-\theta,k)\} \\ &= \sum_{\theta=-\infty}^{\infty} R_{xx}(n-\theta,k)p(\theta) \\ &= \frac{1}{M}\sum_{\theta=0}^{M-1} R_{xx}(n-\theta,k) \\ &= \frac{1}{M}\sum_{m=n}^{M+n-1} R_{xx}(m,k). \end{aligned} \tag{59}$$

Now, observe that

$$\begin{aligned} &\frac{1}{M}\sum_{m=n}^{M+n-1} R_{xx}(m,k) \\ &\quad= \frac{1}{M}\sum_{m=0}^{M-1} R_{xx}(m,k) + \frac{1}{M}\sum_{m=M}^{M+n-1} R_{xx}(m,k) \\ &\qquad- \frac{1}{M}\sum_{m=0}^{n-1} R_{xx}(m,k) \\ &\quad= \frac{1}{M}\sum_{m=0}^{M-1} R_{xx}(m,k) + \frac{1}{M}\sum_{m=M}^{M+n-1} R_{xx}(m,k) \\ &\qquad- \frac{1}{M}\sum_{m=M}^{M+n-1} R_{xx}(m,k) \\ &\quad= \frac{1}{M}\sum_{m=0}^{M-1} R_{xx}(m,k). \end{aligned} \tag{60}$$

The second line follows because $R_{xx}(m,k) = R_{xx}(m+M,k)$ by cyclostationarity. The last sum is independent of n, implying that $R_{\hat{x}\hat{x}}(n,k)$ is a function of k only and that the process $\hat{x}(n)$ is indeed WSS. Furthermore

$$\hat{R}_{xx}(k) = \frac{1}{M}\sum_{n=0}^{M-1} R_{xx}(n,k). \tag{61}$$

APPENDIX C

Proof of (13): Let $x(n)$ be a $(CWSS)_M$ process input to a linear time-invariant filter $P(e^{j\omega})$. The output $z(n)$ is a $(CWSS)_M$ process [10] and is related to $x(n)$ by the convolution sum $z(n) = \Sigma_i\, p(i)x(n-i)$. Our goal is to derive an expression for the average variance of the $(CWSS)_M$ process $z(n)$. Therefore

$$\begin{aligned} \sigma_z^2 &= \frac{1}{M}\sum_{n=0}^{M-1} E\{|z(n)|^2\} \\ &= \frac{1}{M}\sum_{n=0}^{M-1}\sum_i\sum_j p(i)p^*(j)E\{x(n-i)x^*(n-j)\} \\ &= \sum_i\sum_j p(i)p^*(j)\frac{1}{M}\sum_{n=0}^{M-1} R_{xx}(n,j-i) \\ &= \sum_i\sum_j p(i)p^*(j)\hat{R}_{xx}(j-i) \end{aligned} \tag{62}$$

where the last equality follows from (61). By making the change of variables $j-i=l$, we get

$$\begin{aligned} \sigma_z^2 &= \sum_l\sum_j p(j-l)p^*(j)\hat{R}_{xx}(l) \\ &= \sum_l\sum_l r_p(l)\hat{R}_{xx}(l) \end{aligned} \tag{63}$$

where $r_p(l) \triangleq \Sigma_k\, p^*(k)p(k-n)$ is the deterministic autocorrelation of $p(n)$. Taking the discrete-time fourier transform of (63), we get (11). ■

APPENDIX D

Average Power Spectral Density of an Interpolated Random Process: Let $y(n)$ be a wide sense stationary (WSS) random process, input to an interpolation filter, as shown in Fig. 1. The output $x(n)$ is in general a $(CWSS)_M$ process [10]. The average power spectral density of the "stationarized" process has the form

$$\hat{S}_{xx}(e^{j\omega}) = \frac{1}{M}S_{yy}(e^{j\omega M})|F(e^{j\omega})|^2. \tag{64}$$

To derive (64), we can use (61) to write

$$\begin{aligned} \hat{R}_{xx}(k) = \frac{1}{M}\sum_i\sum_j R_{yy}(i-j)\sum_{n=0}^{M-1} f(n-Mi) \\ \cdot f(n-k-Mj). \end{aligned} \tag{65}$$

Making the consecutive change of variables $i-j=l$ and $n-Mi=u$, (65) simplifies to

$$\begin{aligned} \hat{R}_{xx}(k) &= \frac{1}{M}\sum_l R_{yy}(l)\sum_i\sum_{u=-Mi}^{M-1-Mi} f(u)f(u-k+Ml) \\ &= \frac{1}{M}\sum_l R_{yy}(l)\sum_u f(u)f(u-(k-Ml)) \\ &= \frac{1}{M}\sum_l R_{yy}(l)r_f(k-Ml) \end{aligned} \tag{66}$$

where $r_f(n)$ is the deterministic autocorrelation of $f(n)$ as defined in Appendix C. Equation (66) can be interpreted as passing the autocorrelation sequence $(1/M)R_{yy}(n)$ through the interpolation filter $r_f(n)$. Taking the Fourier transform of (66), we obtain (64) or, equivalently, (12). The expression for multiband case (15) can be obtained in a similar fashion.

Again, from (61), we can write

$$\begin{aligned}\hat{R}_{xx}(k) &= \frac{1}{M}\sum_{n=0}^{M-1} E\{x(n)x^*(n-k)\} \\ &= \frac{1}{M}\sum_{n=0}^{M-1}\sum_i\sum_j \boldsymbol{f}(n-Mi)E\{\boldsymbol{y}(n)\boldsymbol{y}^\dagger(n-k)\} \\ &\quad \cdot \boldsymbol{f}^\dagger(n-k-Mj) \\ &= \frac{1}{M}\sum_{n=0}^{M-1}\sum_i\sum_j \boldsymbol{f}(n-Mi)\boldsymbol{R_y}(k) \\ &\quad \cdot \boldsymbol{f}^\dagger(n-k-Mj) \end{aligned} \tag{67}$$

where $\boldsymbol{f}(n) = (f_0(n) \quad f_1(n) \quad \cdots \quad f_{L-1}(n))^T$, and $\boldsymbol{R_y}(k)$ is the autocorrelation matrix of the L WSS inputs $y_k(n)$. By following the same steps used to derive (64), we obtain (13). ■

REFERENCES

[1] A. J. Jerri, "The Shannon sampling theorem—Its various extensions and applications: a tutorial review," *Proc. IEEE*, vol. 65, pp. 1565–1596, Nov. 1977.

[2] G. G. Walter, "A sampling theorem for wavelet subspaces," *IEEE Trans. Inform. Theory*, vol. 38, pp. 881–884, Mar. 1992.

[3] P. P. Vaidyanathan and S.-M. Phoong, "Reconstruction of sequences from non uniform samples," in *Proc. IEEE ISCAS*, Seattle, WA, 1995, vol. 1, pp. 601–604.

[4] ——, "Discrete time signals which can be recovered from samples," in *Proc. IEEE ICASSP*, Detroit, MI, 1995, vol. 2, pp. 1448–1451.

[5] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[6] P. Aziz, H. Sorensen, and J. Van Der Spiegel, "An overview of sigma-delta converters," *IEEE Signal Processing Mag.*, vol. 13, pp. 61–84, Jan. 1996.

[7] J. C. Candy and G. C. Temes, "Oversampling methods for A/D and D/A conversion," in *Oversampling Delta-Sigma Data Converters Theory, Design, and Simulation*. New York: IEEE, 1992.

[8] I. Galton and H. T. Jensen, "Delta-sigma modulator based A/D conversion without oversampling," *IEEE Trans. Circuits Syst. II*, vol. 42, pp. 773–784, Dec. 1995.

[9] E. G. Gladyshev, "Periodically correlated random sequences," *Soviet Math.*, vol. 2, pp. 385–388, 1961.

[10] V. S. Sathe and P. P. Vaidyanathan, "Effects of multirate systems on the statistical properties of random signals," *IEEE Trans. Signal Processing*, vol. 41, pp. 131–146, Jan. 1993.

[11] M. K. Tsatsanis and G. B. Giannakis, "Principal component filter banks for optimal multiresolution analysis," *IEEE Trans. Signal Processing*, vol. 43, pp. 1766–1777, Aug. 1995.

[12] J. Tuqan and P. P. Vaidyanathan, "The role of the discrete-time Kalman-Yakubovich-Popov (KYP) lemma in designing statistically optimum FIR orthonormal filter banks," in *Proc. IEEE ISCAS*, Monterey, CA, June 1998, vol. 5, pp. 122–125.

[13] ——, "A state space approach to the design of globally optimal FIR energy compaction filters," submitted for publication.

[14] P. Moulin and M. K. Mihcak, "Theory and design of signal adapted FIR paraunitary filter banks," *IEEE Trans. Signal Processing*, vol. 46, pp. 920–929, Apr. 1998.

[15] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[16] W. A. Gardner, "Stationarizable random processes," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 8–22, Jan. 1978.

[17] H. L. Hurd, "Stationarizing properties of random shifts," *SIAM J. Appl. Math.*, vol. 26, no. 1, pp. 203–212, Jan. 1974.

[18] W. A. Gardner and L. E. Franks, "Characterization of cyclostationary random processes," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 4–14, Jan. 1975.

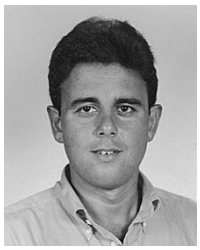
[19] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*. Englewood Cliffs, NJ: Prentice-Hall, 1963.

[20] J. Tuqan and P. P. Vaidyanathan, "Statistically optimum prefiltering and postfiltering in quantization," *IEEE Trans. Circuits Syst. II*, vol. 44, pp. 1015–1031, Dec. 1997.

[21] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Fifth Ed. San Diego, CA: Academic, 1994.

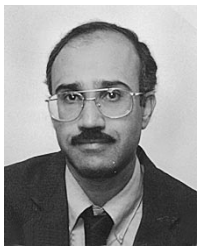
[22] A. Kirac and P. P. Vaidyanathan, "Theory and design of optimum FIR energy compaction filters," *IEEE Trans. Signal Processing*, vol. 46, pp. 903–919, Apr. 1998.

[23] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs: Prentice-Hall, 1989.



Jamal Tuqan (S'91–M'98) was born in Cairo, Egypt, in 1966. He received the B.S. degree (with honors) in electrical engineering from Cairo University in 1989, the M.S.E.E. degree from the Georgia Institute of Technology, Atlanta, in 1992, and the Ph.D. degree in electrical engineering from the California Institute of Technology (CalTech), Pasadena, in 1997.

From September 1989 to July 1990, he worked at the IBM Research Center, Cairo, writing software for Arabic speech compression based on linear predictive coding techniques. From July 1994 to December 1997, he was a Research and Teaching Assistant at CalTech. From January to June 1998, he was a Technical Staff Member in the Digital Signal Processing Group at the same university. During the spring quarter 1998, he was appointed as a Lecturer at Caltech to teach a class on linear estimation theory and adaptive filters. He is currently with the Image and Video Communications Group at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY. His main research interests are in the general area of digital signal processing, signal processing applications in communications, data compression, and applied mathematics. The focus of his Ph.D. thesis research was on the optimization of multirate systems and filter banks according to the input signal statistics for compression and communication applications.



P. P. Vaidyanathan (S'80–M'83–SM'88–F'91) was born in Calcutta, India, on October 16, 1954. He received the B.Sc. (Hons.) degree in physics and the B.Tech. and M.Tech. degrees in radiophysics and electronics, all from the University of Calcutta, in 1974, 1977, and 1979, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 1982.

He was a Post Doctoral Fellow at the University of California, Santa Barbara, from September 1982 to March 1983. In March 1983, he joined the Department of Electrical Engineering, California Institute of Technology, Pasadena, as an Assistant Professor, and since 1993, he has been Professor of Electrical Engineering. His main research interests are in digital signal processing, multirate systems, wavelet transforms, and adaptive filtering.

Dr. Vaidyanathan served as Vice Chairman of the Technical Program Committee for the 1983 IEEE International Symposium on Circuits and Systems and as the Technical Program Chairman for the 1992 IEEE International Symposium on Circuits and Systems. He was an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS from 1985 to 1987 and is currently an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS and a consulting editor for the journal *Applied and Computational Harmonic Analysis*. He has been a Guest Editor in 1998 for Special Issues of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the IEEE TRANSACTIONS ON CICUITS AND SYSTEMS II on the topics of filter banks, wavelets, and subband coders. He has authored a number of papers in IEEE journals and is the author of the book *Multirate Systems and Filter Banks* (Englewood Cliffs, NJ: Prentice-Hall, 1993). He has written several chapters for various signal processing handbooks. He was a recepient of the Award for Excellence in Teaching at the California Institute of Technology for the years 1983 to 1984, 1992 to 1993, and 1993 to 1994. He also received the NSF's Presidential Young Investigator award in 1986. In 1989, he received the IEEE ASSP Senior Award for his paper on multirate perfect-reconstruction filter banks. In 1990, he was recepient of the S. K. Mitra Memorial Award from the Institute of Electronics and Telecommunications Engineers, India, for his joint paper in the IETE journal. He was also the coauthor of a paper on linear-phase perfect reconstruction filter banks in the IEEE TRANSACTIONS ON SIGNAL PROCESSING, for which the first author (T. Q. Nguyen) received the Young Outstanding Author Award in 1993. He received the 1995 F. E. Terman Award of the American Society for Engineering Education, sponsored by Hewlett Packard Co., for his contributions to engineering education, especially the book *Multirate Systems and Filter Banks*. He was a distinguished lecturer for the IEEE Signal Processing Society for the year 1996 to 1997.